HHAI 2024:
HYBRID HUMAN AI SYSTEMS FOR THE SOCIAL GOOD
Proceedings of the Third International Conference on Hybrid Human-Artificial Intelligence

Edited by
Fabian Lorig
Jason Tucker
Adam Dahlgren Lindström
Frank Dignum
Pradeep Murukannaiah
Andreas Theodorou
Pınar Yolum
HHAI 2024: HYBRID HUMAN AI SYSTEMS FOR THE SOCIAL GOOD
Frontiers in Artificial Intelligence and Applications

The book series Frontiers in Artificial Intelligence and Applications (FAIA) covers all aspects of theoretical and applied Artificial Intelligence research in the form of monographs, selected doctoral dissertations, handbooks and proceedings volumes. The FAIA series contains several sub-series, including ‘Information Modelling and Knowledge Bases’ and ‘Knowledge-Based Intelligent Engineering Systems’. It also includes the biennial European Conference on Artificial Intelligence (ECAI) proceedings volumes, and other EurAI (European Association for Artificial Intelligence, formerly ECCAI) sponsored publications. The series has become a highly visible platform for the publication and dissemination of original research in this field. Volumes are selected for inclusion by an international editorial board of well-known scholars in the field of AI. All contributions to the volumes in the series have been peer reviewed. The FAIA series is indexed in ACM Digital Library; DBLP; EI Compendex; Google Scholar; Scopus; Web of Science: Conference Proceedings Citation Index – Science (CPCI-S) and Book Citation Index – Science (BKCI-S); Zentralblatt MATH.

Series Editors:
Nicola Guarino, Pascal Hitzler, Joost N. Kok, Jiming Liu, Ramon López de Mántaras, Riichiro Mizoguchi, Mark Musen, Sankar K. Pal, Ning Zhong

Volume 386

Recently published in this series

Vol. 385. Y. Ye and P. Siarry (Eds.), Artificial Intelligence and Human-Computer Interaction – Proceedings of the 1st International Conference (ArtInHCI 2023), Wuhan, China, 27–28 October 2023
Vol. 383. L.C. Jain, V.E. Balas, Q. Wu and F. Shi (Eds.), Design Studies and Intelligence Engineering – Proceedings of DSIE 2023, Hangzhou, China, 28–29 October 2023
Vol. 379. G. Sileno, J. Spanakis and G. van Dijck (Eds.), Legal Knowledge and Information Systems – JURIX 2023: The Thirty-sixth Annual Conference, Maastricht, the Netherlands, 18–20 December 2023

ISSN 0922-6389 (print)
ISSN 1879-8314 (online)
Preface

This volume presents the proceedings of the 3rd International Conference on Hybrid Human-Artificial Intelligence (HHAI 2024), held in Malmö, Sweden, from 10–14 June 2024. The focus of HHAI 2024 was on artificially-intelligent systems that cooperate synergistically, proactively, and purposefully with humans, amplifying rather than replacing human intelligence.

The HHAI field is driven by developments in AI, but it also requires fundamentally new approaches and solutions. For this reason, we encourage collaboration across research domains such as AI, HCI, cognitive and social sciences, philosophy & ethics, complex systems, and others. For this third international conference, we invited scholars from these fields to submit their best, original – new as well as in progress – work, and visionary ideas on hybrid human-artificial intelligence.

The following list of topics is illustrative, not exhaustive:

- Human-AI interaction and collaboration
- Adaptive human-AI co-learning and co-creation
- Learning, reasoning and planning with humans and machines in the loop
- User modelling and personalisation
- Integration of learning and reasoning
- Transparent, explainable, and accountable AI
- Fair, ethical, responsible, and trustworthy AI
- Societal awareness of AI
- Multimodal machine perception of real-world settings
- Social signal processing
- Representations learning for communicative or collaborative AI
- Symbolic and narrative-based representations for human-centric AI
- The role of design and compositionality of AI systems in interpretable/collaborative AI

Contributions about all types of technology, from robots and conversational agents to multi-agent systems and machine learning models were welcome.

Acknowledgments

This edition of Hybrid Human-Artificial Intelligence was organised by the Hybrid Intelligence\footnote{https://www.hybrid-intelligence-centre.nl} and Humane-AI European Network, which also contributed financially to the conference, and was supported by The Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS\footnote{https://wasp-hs.org}), AI Policy Lab\footnote{https://aipolicylab.se} and hosted by Malmö University, with support from Umeå University. Esra Karabiber and the Malmö University Conference Service also provided invaluable support in the organising of the conference.
We would like to take this opportunity to thank everybody who submitted their work for review and all those who presented their work at the conference. Special thanks also to the members of the programme committee, the organisers of the pre-conference workshops, tutorials, and creative events, and the sponsors of the conference for their contributions.

Fabian Lorig (General Chair)
Jason Tucker (General Chair)
Adam Dahlgren Lindström (General Chair)
Frank Dignum (General Chair)
Pradeep Murukannaiah (Program Chair)
Andreas Theodorou (Program Chair)
Pınar Yolum (Program Chair)
About the Conference

Organising Committee

Conference Chairs
Fabian Lorig (Malmö University, Sweden)
Jason Tucker (Malmö University, Sweden)
Adam Dahlgren Lindström (Umeå University, Sweden)
Frank Dignum (Umeå University, Sweden)

Programme Chairs
Pradeep Murukannaiah (TU Delft, The Netherlands)
Andreas Theodorou (Universitat Politècnica de Catalunya, Spain)
Pınar Yolum (Utrecht University, The Netherlands)

Workshop and Tutorial Chairs
Petter Ericsson, (Umeå University, Sweden)
Nina Khairova (Umeå University, Sweden)
Marina de Vos (University of Bath, United Kingdom)

Demo and Poster Chairs
Michael Belfrage (Malmö University, Sweden)
Cezara Pastrav (Umeå University, Sweden)

Doctoral Consortium Chairs
Passant El.Agroudy (DFKI, Germany)
Michiel van der Meer (Leiden University, The Netherlands)
Harko Verhagen (Stockholm University, Sweden)

Diversity and Inclusion Chairs
Emil Johansson (Malmö University, Sweden)
Annet Onnes (Utrecht University, The Netherlands)

Satellite Events and Industry Chair
Sandor Albrecht (Knut och Alice Wallenbergs Foundation, Sweden)

Publicity and Social Media Chairs
Jesse Grootjen (LMU Munich, Germany)
Julian Rasch (LMU Munich, Germany)

Student Volunteer Chairs
Andreas Brännström (Umeå University, Sweden)
Leila Methnani (Umeå University, Sweden)
Finance Chair
Tatyana Sarayeva (Umeå University, Sweden)

Web Chair
Bertilla Fabris (Malmö University, Sweden)

Local Support
Esra Karabiber (Malmö University, Sweden)

Programme Committee
Erman Acar (University of Amsterdam)
Muneeb Ahmad (Western Sydney University)
Nirav Ajmeri (University of Bristol)
Rachid Alami (LAAS-CNRS (ANITI (University of Toulouse)
Huib Aldewereld (HU University of Applied Sciences Utrecht)
Hüseyin Aydın (Middle East Technical University (Utrecht University)
Daniel Balliet (Vrije Universiteit)
Michael Belfrage (Malmö University)
Victor de Boer (Vrije Universiteit Amsterdam)
Harmen de Weerd (Hanz University of Applied Sciences)
Chiara Boldrini (CNR-IIT)
Joost Broekens (LIACS)
Andreas Brännström (Umeå University)
Sahan Bulathwela (University College London)
Daniel Buschek (University of Bayreuth)
Rafael A. Calvo (Imperial College London)
Sarah E. Carter (TU Delft)
Luciano Cavalcante Siebert (Delft University of Technology)
Mustafa Celikok (TU Delft)
Manolis Chiou (Queen Mary University of London)
Mathieu D'Aquin (LORIA (University of Lorraine)
Marina De Vos (University of Bath)
Davide Dell'Anna (Utrecht University)
Tilman Dingler (Delft University of Technology)
Lucia Donatelli (Vrije Universiteit Amsterdam)
Philipp Ebel (University of St. Gallen)
Passant Elagroudy (DFKI, German Research Center for Artificial Intelligence)
Zeynep Engin (Data for Policy CIC)
Petter Ericson (Umeå University)
Thommy Eriksson (Chalmers University of Technology)
Kristian Fenech (Eötvös Loránd University)
Antske Fokkens (Vrije Universiteit Amsterdam)
Mary Ellen Foster (University of Glasgow)
Swen Gaudl (University of Gothenburg)
Ben Green (University of Michigan)
Marko Grobelnik (Jožef Stefan Institute (Ljubljana)
Davide Grossi (University of Groningen)
Conference Programme and Contributions

Following the examples of HHAI 2022 and 2023, HHAI 2024 was organised as a single-track programme. This was based on the positive experience of adopting this structure in previous years, a practice which has facilitated the ongoing exchange and development of ideas across a range of fields.

The main track of the conference accepted three types of paper, 1) full papers presenting original, impactful work; 2) blue-sky papers presenting visionary ideas to stimulate the research community; and 3) working papers presenting work in progress. Unlike previous iterations, HHAI 2024 followed a double-blind review process.

There were 62 submissions for the main track, of which 31 were accepted (9 full papers, 5 blue sky, and 17 working papers) and 31 rejected. The final acceptance rate for full papers was 29% (50% across all tracks of the main programme).

This volume contains all submissions accepted for the main track – including full papers, blue sky papers and working papers – proposals from the Doctoral Consortium, and extended abstracts from the Posters and Demos track.

The keynote speakers at the main conferences were as follows: Dr. Fosca Giannotti, Professor at Scuola Normale Superiore, Pisa & Information Science and Technology Institute “A. Faedo” of CNR., Pisa, Italy; Dr. Mark Riedl, professor in the Georgia Tech School of Interactive Computing and associate director of the Georgia Tech Machine Learning Center; and Dr. Anastasia Varava, Research Lead at SEBx.

Additional Programme

In addition to the main conference, HHAI 2024 also hosted a diverse set of workshops and tutorials on 10 and 11 June, as well as a Doctoral Consortium and the first HHAI Summer School. New this year were two discovery workshops providing an informal way to gather around interesting topics and establish starting points for further collaboration.

There were 17 proposals for workshops, tutorials, and creative events. The tutorial Introduction to Human Hybrid AI was added to complement the programme and facilitate the summer school students. There were 20 submissions to the Posters and Demos track as well as 12 submissions to the Doctoral Consortium.

Tutorials

- Introduction to Human Hybrid AI - Organisers: John Shawe-Taylor, Frank van Harmelen, Virginia Dignum, Frank Dignum
- Human-AI mutual promotion for emotion and cognition understanding - Organisers: Haoyu Chen, Andy Nguyen, Yang Liu and Sanna Järvelä
- Collaborative HAI-Learning through Conceptual Exploration - Organisers: Bernhard Ganter, Tom Hanika, Johannes Hirth and Sergei Obiedkov
- Tutorial on Contextualising and Executing Robot Manipulation Plans Using Web Knowledge - Organisers: Michael Beetz, Philipp Cimiano, Michaela Kümpel, Jan-Philipp Töber and Ilaria Tiddi, Enrico Motta
- *An Introduction to Computational Argumentation* - Organisers: Elfia Bezou-Vrakatseli, Daphne Odekerken and Andreas Xydis
- *Interactive Concept-based Search* - Organiser: Amiram Moshaiov
- *Knowledge Engineering for Hybrid Intelligence (KE4HI)* - Organisers: Ilaria Tiddi, Victor de Boer and Stefan Schlobach

**Creative Event**

- *The What-is-HI Competition* - Organiser(s): Davide Dell’Anna, Bernd Dudzik, Davide Grossi, Catholijn Jonker, Pradeep Kumar Murukannaiah and Pinar Yolum

**Workshops**

- *Exploring Tangible AI with Theory Instruments* - Organisers: Jacob Buur, Mette Gislev Kjersgaard, Ona Pirol and Jessica Sorensen.
- *Workshop on Responsible Applied Artificial Intelligence (RAAIIT)* - Organisers: Maaikie Harbers, Stefan Leijen, Pascal Wiggers, Marieke Peeters, Saskia Robben, Roland van Dierendonck, Fabian Kok, Sophie Horsman and Tiwaaaee van der Horst.
- *Hybrid Intelligence for Health Care* - Organisers: Chenxu Hao, Mark Neerincx, Myrthe Tielman, Jasper van der Waa and Maaikie de Boer
- *Stimulating cognitive engagement in hybrid decision-making: friction, reliance and biases* - Organisers: Chiara Natali, Brett Frischmann and Federico Cabitza
- *Imagine the AI Landscape after the AI Act* - Organisers: Francesca Pratesi, Desara Dushi and Francesca Naretto.
- *AI in Africa and SDGs: Bridging Networks and Fostering Climate Action* - Organisers: John Shawe-Taylor, Davor Orlic and Essa Mohamedali

The first edition of the HHAI Summer Schools will take place as part of HHAI 2024. Students were invited to participate in a series of dedicated events exploring the cooperation between humans and intelligent systems. As part of the pre-conference programme, leading researchers and experts from industry offered a series of tutorials, workshops, and other events covering cutting-edge topics related to Hybrid Human-
Artificial Intelligence. To receive a certificate of attendance, students were expected to attend the following sessions:

- Monday – Tuesday: four tutorials or three tutorials and the Doctoral Consortium
- Wednesday – Friday: the discovery workshops and keynotes
- Attending the other session of the main conference was advised

The Summer School was supported by HumanE-AI-Net¹, a network of European research centres, universities, and key industrial champions that combines world leading AI competence with key players in related areas such as HCI, cognitive science, social sciences and complexity science. HumanE-AI-Net’s goal is to lay the foundations of designing the principles for a new science that will produce AI based on European values and with more relevance for Europeans.

Beyond the main programme, HHAI 2024 organised several social events in the evening throughout the conference, including a reception on 11 June and Gala Dinner on 12 June.

¹ https://www.humane-ai.eu
ORGANISED BY

HUMANE AI NET

SUPPORTED BY

Hybrid Intelligence

WASP-HS

AI Policy Lab
at Umeå University
This page intentionally left blank
Contents

Preface v

Fabian Lorig, Jason Tucker, Adam Dahlgren Lindström, Frank Dignum, Pradeep Murukannaiah, Andreas Theodorou and Pınar Yolum

About the Conference vii

Main Track Papers

LLM-Generated Word Association Norms 3
Katherine Abramski, Clara Lavorati, Giulio Rossetti and Massimo Stella

Beyond Skills: The Role of Values in Job Seeking in the Era of Industry 5.0 13
Ajaya Adhikari, Quirine Smit, Joost A.G.M. van Genabeek, Steven Dhondt and Cor J. Veenman

Formal Specification of Actual Trust in Multiagent Systems 22
Michael Akintunde, Vahid Yazdanpanah, Asieh Salehi Fathabadi, Corina Cirstea, Mehdi Dastani and Luc Moreau

In Search of Clarity: Discerning Between Human Replacement and Augmentation 36
Costanza Alfieri, Martina De Sanctis, Donatella Donati and Paola Inverardi

Playing the Imitation Game: Human-AI Simulators in Pedagogic Design 46
Florentina Armaselu

Assisting Users in Privacy Conflicts with Partially Observable Multi-Agent Reinforcement Learning 55
Hüseyin Aydın

Exploring Large Language Models Capabilities to Explain Decision Trees 64
Paulo Bruno Serafim, Pierluigi Crescenzi, Gizem Gezici, Eleonora Cappuccio, Salvatore Rinzivillo and Fosca Giannotti

An Interactive Interface for Feature Space Navigation 73
Eleonora Cappuccio, Isacco Beretta, Marta Marchiori Manerba and Salvatore Rinzivillo

Human-Centered AI for Dementia Care: Using Reinforcement Learning for Personalized Interventions Support in Eating and Drinking Scenarios 84
Wen-Tseng Chang, Shihan Wang, Stephanie Kramer, Michel Oey and Somaya Ben Allouch

Hybrid Intelligence for Digital Humanities 94
Victor de Boer and Lise Stork

EASY-AI: sEmantic And compoSable glYphs for representing AI systems 105
Alexis Ellis, Brandon Dave, Hugh Salehi, Subhashini Ganapathy and Cogan Shimizu
Effective Maintenance of Computational Theory of Mind for Human-AI Collaboration
  Emre Erdogan, Frank Dignum and Rineke Verbrugge 114

Towards Effective Management of Verbal Probability Expressions Using a Co-Learning Approach
  Christian Fleiner and Joost Vennekens 124

LLM-Augmented Agent-Based Modelling for Social Simulations: Challenges and Opportunities
  Önder Gürcan 134

Investigating Labeler Bias in Face Annotation for Machine Learning
  Luke Haliburton, Sinkas Ghebremedhin, Robin Welsch, Albrecht Schmidt and Sven Mayer 145

Human-AI Collaboration in Smart Manufacturing: Key Concepts and Framework for Design
  Maria Hartikainen, Guna Spurava and Kaisa Väänänen 162

Assessing the HI-ness of Virtual Heritage Applications with Knowledge Engineering
  Delaram Javdani Rikhtehgar, Ilaria Tiddi, Shenghui Wang, Stefan Schlobach and Dirk Heylen 173

Evaluating Generative AI Incidents: An Exploratory Vignette Study on the Role of Trust, Attitude and AI Literacy
  Esther S. Kox and Beatrice Beretta 188

Unreflected Acceptance – Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education
  Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann and Jakob Karolus 199

“Computer, Generate!” – Investigating User-Controlled Generation of Immersive Virtual Environments
  Carina Liebers, Niklas Pfützenreuter, Jonas Auda, Uwe Gruenefeld and Stefan Schneegass 213

Human-AI Engineering for Adults
  André Meyer-Vitali and Wico Mulder 228

Hybrid Intelligence in Academic Writing: Examining Self-Regulated Learning Patterns in an AI-Assisted Writing Task
  Andy Nguyen, Faith Ilesanmi, Belle Dang, Eija Vuorenmaa and Sanna Järvelä 241

The Eve of AI Reclaiming Womanhood Through Co-Design with Artificial Intelligence
  Sangeetha Ramanath 255
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust in AI Chatbots: The Perceived Expertise of ChatGPT in Subjective and Objective Tasks</td>
<td>264</td>
</tr>
<tr>
<td>Maike Ramrath, Alisa Scharmann, Annika Ridder, Tristan Kühn, Sophie Weller and Nicole Krämer</td>
<td></td>
</tr>
<tr>
<td>Common Ground Provides a Mental Shortcut in Agent-Agent Interaction</td>
<td>281</td>
</tr>
<tr>
<td>Ramira van der Meulen, Rineke Verbrugge and Max van Duijn</td>
<td></td>
</tr>
<tr>
<td>Building a Stronger Case: Combining Evidence and Law in Scenario-Based Bayesian Networks</td>
<td>291</td>
</tr>
<tr>
<td>Ludi van Leeuwen, Rineke Verbrugge, Bart Verheij and Silja Renooij</td>
<td></td>
</tr>
<tr>
<td>The Influence of Interdependence on Trust Calibration in Human-Machine Teams</td>
<td>300</td>
</tr>
<tr>
<td>Ruben S. Verhagen, Alexandra Marcu, Mark A. Neerincx and Myrthe L. Tielman</td>
<td></td>
</tr>
<tr>
<td>A Conversational Agent for Structured Diary Construction Enabling Monitoring of Functioning &amp; Well-Being</td>
<td>315</td>
</tr>
<tr>
<td>Piek Vossen, Selene Báez Santamaria and Thomas Baier</td>
<td></td>
</tr>
<tr>
<td>Human-Aware Planning for Situational Awareness in Indoor Police Interventions</td>
<td>325</td>
</tr>
<tr>
<td>Joost Vossers, Andreas Brännström, Erik Borglund, Jonas Hansson and Juan Carlos Nieves</td>
<td></td>
</tr>
<tr>
<td>Exploring the Dynamic Nature of Trust Using Interventions in a Human-AI Collaborative Task</td>
<td>335</td>
</tr>
<tr>
<td>Sachini Weerawardhana, Michael Akintunde and Luc Moreau</td>
<td></td>
</tr>
<tr>
<td>Using Default Logic to Create Adaptable User Models for Behavior Support Agents</td>
<td>350</td>
</tr>
<tr>
<td>Johanna Wolff, Victor de Boer, Dirk Heylen and M. Birna van Riemsdijk</td>
<td></td>
</tr>
</tbody>
</table>

**Doctoral Consortium Papers**

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploring Human-AI Collaboration and Explainability for Sustainable ML</td>
<td>363</td>
</tr>
<tr>
<td>Eya Ben Chaaben</td>
<td></td>
</tr>
<tr>
<td>Investigations to Facilitate Human-Centred AI Development in Manufacturing</td>
<td>371</td>
</tr>
<tr>
<td>Maria Hartikainen</td>
<td></td>
</tr>
<tr>
<td>Developing Meaningful Explanations for Machine Learning Models in the Telecom Domain</td>
<td>379</td>
</tr>
<tr>
<td>Henry Maathuis</td>
<td></td>
</tr>
<tr>
<td>Towards Value Alignment for Opaque Agents Through Concept Analysis and Inter-Agent Value Modelling</td>
<td>386</td>
</tr>
<tr>
<td>Jack McKinlay</td>
<td></td>
</tr>
<tr>
<td>Neuro-Symbolic Approach for Tantrum Monitoring and Prevention in Individuals with Autism Spectrum Disorder: A Protocol for Virtual Agents</td>
<td>394</td>
</tr>
<tr>
<td>Andrea Monaldini</td>
<td></td>
</tr>
</tbody>
</table>
Explainable Interactive Machine Learning Using Prototypical Part Networks for Medical Image Analysis  
\textit{Alec Parise and Brian Mac Namee}  
402

Designing Expressive Interaction with Generative Artificial Intelligence  
\textit{Xiaohan Peng}  
409

Interactive Multi-Objective Optimization for Military Helicopter Route Planning  
\textit{Thomas Quadt}  
418

Artificial Intelligence in Wearables – Challenges and Opportunities in Physical Therapy and Sports Training  
\textit{Joanna Sorysz}  
427

Integrating Minds and Machines: The Role of Digital Externalization in Evolving Cognitive Architectures  
\textit{Gleb Vzorin}  
434

Hypothesis Updating by Combining Knowledge Graphs and Argumentation  
\textit{Giacomo Zamprogno}  
441

\textbf{Posters/Demos}

Cognitive Network Science Unveils Affective Bias in GPT Models Mirroring Math Anxiety in High-School Students  
\textit{Katherine Abramski, Salvatore Citraro, Luigi Lombardi, Giulio Rossetti and Massimo Stella}  
451

Why Do You Think This Person Is Introverted? – Towards Conversational Argumentative Explainability  
\textit{Alexander Berman}  
454

\textit{Pei-Yu Chen, Selene Baez Santamaria, Maaike H.T. de Boer, Floris den Hengst, Bart A. Kamphorst, Quirine Smit, Shihan Wang and Johanna Wolff}  
457

Institutional Modelling: A Case Study of the Swedish Organ Donation System  
\textit{Bertilla Fabris, Michael Belfrage and Fabian Lorig}  
460

Keep Gesturing: A Game for Pragmatic Communication  
\textit{János Adrián Gulyás, Miklós Máté Badó, Kristian Fenech and András Lörincz}  
463

What Happens to a Robotic Startup While It Evolves  
\textit{Alexandra Hettich}  
466

The Impact of Mixed-Initiative on Collaboration in Hybrid AI  
\textit{Leila Methnani, Adam Dahlgren Lindström and Virginia Dignum}  
469

A Hybrid Intelligent Change Management Approach to Generative AI Adoption  
\textit{Jacob Sherson}  
472
Operational Criteria of Hybrid Intelligence for Generative AI Virtual Assistants
Jacob Sherson, Janet Rafner and Safinaz Büyükgüzel

Facilitating Human Feedback for GenAI Prompt Optimization
Jacob Sherson and Florent Vinchon

Value-Sensitive Disagreement Analysis for Online Deliberation
Michiel van der Meer, Piek Vossen, Catholijn M. Jonker and Pradeep K. Murukannaiah

ChatGPT May Foster Human Gist Memory While Offloading Less Sufficient Information
Gleb Vzorin, Vasilissa Petrova and Anna Sedykh

AI + Object Making & Creative Content
Kio Zhu and William Nickley

Author Index
Main Track Papers
This page intentionally left blank
LLM-Generated Word Association Norms

Katherine ABRAMSKI, Clara LAVORATI, Giulio ROSSETTI and Massimo STELLA

Dept. of Computer Science, University of Pisa, Italy
Institute of Information Science and Technologies, CNR, Italy
Dept. of Psychology and Cognitive Science, University of Trento, Italy
these authors contributed equally

Abstract. Word associations have been extensively used in psychology to study the rich structure of human conceptual knowledge. Recently, the study of word associations has been extended to investigating the knowledge encoded in LLMs. However, because of how the LLM word associations are accessed, existing approaches have been limited in the types of comparisons that can be made between humans and LLMs. To overcome this, we create LLM-generated word association norms modeled after the Small World of Words (SWOW) human-generated word association norms consisting of over 12,000 cue words. We prompt the language models with the same cues and participant profiles as those in the SWOW human-generated norms, and we conduct preliminary comparative analyses between humans and LLMs that explore differences in response variability, biases, concreteness effects, and network properties. Our exploration provides insights into how LLM-generated word associations can be used to investigate similarities and differences in how humans and LLMs process information.

Keywords. artificial intelligence, large language models, network science, word associations, mental lexicon, cognitive modeling, cognitive bias

1. Introduction

Understanding the mental lexicon is an important part of studying human cognition. The way in which we represent concepts in terms of relationships between them is essential for understanding how we process information, and consequently, how we reason, form beliefs, and make decisions [1]. One way to tap into the rich structure of human conceptual knowledge is through word associations, obtained by asking participants to produce associated responses when prompted with a cue word. Word associations have been extensively used in cognitive psychology and linguistics for studying lexical retrieval [2,3], semantic organization [4], and similarity judgments [5,6,7], as well as investigating concreteness effects [8,9] and cognitive biases [10]. Additionally, cognitive models built from word association norms can provide powerful insights into many different cognitive phenomena, such as language learning [11] and creativity [12].
Recently, researchers have begun to use word associations as a method for investigating the capabilities and limitations of LLMs [13], including investigations of biases [14,15]. Most of these approaches investigate the embedding space of language models in order to gain access to their associations [16]. They then compare these LLM associations to the well studied properties of human-generated word associations, such as asymmetry and intransitivity [17]. While this approach provides important insights, it is limited for a few reasons. First, since associations extracted from embedding spaces are usually based on cosine similarity, they are symmetrical, unlike human associations. Also, contextual embeddings must first be transformed into static embeddings [18], which can introduce bias and distort similarity estimates [19]. Challenges such as these related to investigating contextual embeddings have led to a broader shift in how researchers approach investigating LLMs [19], from a bottom-up approach to a top-down approach that probes LLMs in a variety of cognitive and linguistic tasks in order to better understand their capabilities and reconstruct their cognitive architecture [20,21,22]. This new approach has led to the emergence of machine psychology [23], a new field which entails applying the tools of cognitive psychology to investigate the behavior of machines as if they were human participants in psychological experiments. The machine psychology approach has several advantages. First, probing methods can be applied regardless of the type of LLM – a significant advantage considering the rapid pace of LLM advancements – and also, the top-down approach allows for more direct comparisons between humans and LLMs.

One recent study applied a machine psychology approach to compare human word associations with LLM word associations accessed from rule mining on word sequences sampled from LLMs [17]. While this approach more closely imitates human word associations, the probing method still differs significantly from how human word associations are accessed, limiting the types of comparisons that can be made. In this work, we aim to close that gap by creating datasets of LLM-generated word association norms that are directly comparable to human-generated norms. We model our dataset after the Small World of Worlds English word association norms (SWOW) [7], the largest and most recent dataset to date. We prompt Mistral AI’s large language model – in particular the mistral-7b-instruct-0.1 model – to produce responses to the same exact cues that are present in the SWOW dataset. We create two datasets of LLM-generated norms: for the first dataset, we prompt the model with only the cue words. For the second, we prompt the model with cue words as well as the exact profiles of the participants in the original SWOW experiment (i.e. age, gender, etc.). In this way, we also investigate how well the model is capable of simulating (its own interpretation of) a specific profile.

The aim of this working paper is to provide an overview of the datasets and a brief demonstration of how they may be used. The remainder of the paper is organized as follows. In Section 2 we present the methodology used to generate and preprocess the data and we discuss the preliminary comparative analyses that we performed on all three sets of norms: the original SWOW dataset, the Mistral-without-profiles dataset, and the Mistral-with-profiles dataset. In Section 3 we present the results of our preliminary analyses, and in Section 4 we briefly discuss directions for future work.
2. Experimental settings and methodology

We prepared the input to the model using the preprocessed original SWOW dataset containing 12,282 unique cue words, each repeated 100 times, with three responses (R1, R2, and R3) per cue token. We matched the SWOW participant profiles to their corresponding cue words to ensure that the Mistral-with-profiles dataset would be aligned with the original SWOW data. We then prompted the model to provide three associations each time it was presented with a cue word. In the case of the dataset with participant profiles, we also asked the model to respond as if it were a person with the specified profile.

Preprocessing of the responses consisted of various steps. First, proper names and spelling errors were corrected (including changing British spelling to American spelling) using mapping tables from the original SWOW experiment. Prefixes the, a an, and and were also removed from the responses, unless the response was among the cues (e.g. a lot). Then, a series of ad-hoc filters were applied to remove nonsensical responses such as printassociation, corresponding to 0.75% and 4.32% of all responses in the Mistral-without-profiles data and the Mistral-with-profiles data, respectively. Finally, duplicate responses and responses identical to their cues were removed.

Following data preprocessing, we performed the following preliminary exploratory analyses to capture peculiarities of the generated datasets. Properties of cues and responses. We investigated the properties of cues and responses by counting the numbers of tokens, numbers of types, and the percentage of missing responses in all three datasets. We also calculated the percentage overlap of the original responses compared to the LLM responses. Additionally, we compared the distributions of the number of unique responses per cue across the three datasets. These statistics provide important insights about the richness of the responses provided by humans compared to LLMs.

Investigating biases and relations. Since word associations are generally spontaneous and automatic, they can serve as a window into our implicit biases. In order to explore any potential differences in gender biases across the datasets, we looked at the top ten most frequent response tokens to the cues man and woman. We also investigated differences in the types of relations that responses shared with the cues, specifically, paradigmatic vs. syntagmatic relations. Paradigmatic relations are those that can be expressed in a taxonomy or can be substituted for each other [24], including synonymy (i.e. woman – lady), antonymy (i.e. woman – man), and hypernymy (i.e. woman – person). Rather, syntagmatic relations are those that tend to occur in similar contexts [25](i.e. woman – feminism). Whether a response has a paradigmatic or a syntagmatic relation with the cue word has important implications for how lexical data is processed, and may also provide insights about how we form biases.

Concreteness effects. Concreteness effects are nuanced and complex differences in how we process lexical information with regards to the abstractness/concreteness

---

2The link to the repository containing the LLM-generated norms will be made available upon the acceptance of the working paper.
of a word. We were interested in investigating the concreteness effect that concrete words have stronger but fewer associates while abstract words have weaker but more associates [8]. We investigated this effect in all three datasets.

**Network comparisons.** Representing word association norms as complex networks enables us to investigate structural properties of the mental lexicon that would otherwise not emerge. We built weighted directed networks from the three datasets such that cues are source nodes and responses are target nodes. Therefore, edges are directed from cues to responses and weighted based on the frequency of the association. We then considered only the largest strongly connected components, keeping only those nodes that were both cues and responses. We report network statistics for these three networks (i.e., density, clustering coefficient, etc.) and then we make pairwise comparisons of the three networks to quantify how similar and different they are. First, we made pairwise comparisons of their sets of nodes. We use the Jaccard coefficient (i.e., \((A \cap B)/(A \cup B))\) as a measure of similarity, and we also calculated the respective set difference percentages (i.e., \((A - B)/A\) and \((B - A)/B\)). To compare sets of edges, we considered only the edges in the node intersection of the networks being compared. We then calculated the same measures that we used to assess similarities and differences between the sets of nodes, that is, the Jaccard coefficient and the respective set difference percentages.

### 3. Experimental Results

In this section we report the preliminary results obtained from performing the exploratory analyses defined in Section 2.

**Properties of cues and responses** Table 1 displays the statistics reflecting the properties of the cues and responses of the three datasets. All three datasets have the same cues, and they differ only in their responses. The Mistral-without-profiles norms have the lowest percentage of missing responses, while the Mistral-with-profiles norms have the highest percentage of missing responses. The number of response types (unique responses), however, is significantly higher in the original dataset compared to the LLM datasets. The Mistral-without-profiles dataset also has slightly more response tokens than the Mistral-with-profiles dataset. In line with these statistics, 77.3% of the original response types are not in the Mistral-without-profiles responses, compared to just 30.7% of the Mistral-without-profiles response types that are not in the original responses. These percentages diverge even further to 82.9% and 29.9% respectively when comparing the original data to the Mistral-with-profiles data. These statistics indicate that humans generate a much wider variety of responses than both the LLMs, and that Mistral-without-profiles generates a slightly wider variety of responses compared to Mistral-with-profiles. These statistics are also reflected in the histograms in Figure 1, that display the number of unique responses per cue for the three datasets. Unlike the human distribution, the LLM ones are skewed right, reflecting the tendency to produce fewer unique responses to cues compared to humans.

**Investigating biases and relations.** Table 2 displays the top ten most frequent response tokens to *man* and *woman*. We immediately notice that the responses pro-
Table 1. Statistics for cues, R1, R2, R3, and R123 (all responses combined) are provided for the three datasets. Statistics include the number of tokens (total counts), the number types (unique counts), and the percentage of missing response tokens.

![Figure 1. Histograms displaying the number of unique responses per cue for the original data (left), Mistral data without profiles (center), and Mistral data with profiles. While the original data are quite symmetrical, the Mistral data are skewed right, tending towards fewer unique responses.](image)

Table 2. The top ten most frequent responses to the cues man and woman for the three datasets are shown. Responses shown in bold reflect paradigmatic relations with the cue word, while responses shown in italics reflect syntagmatic relations with the cue word.

duced by the LLMs are blatant stereotypical gender biases (e.g. woman – makeup, man – career). While we observe some biases among the human-generated responses (e.g. man – strong), they are not nearly as pronounced as those pro-
duced by the LLMs. We also notice that responses produced by humans tend towards paradigmatic relations with the cue words (shown in bold in Table 3) while responses produced by LLMs tend towards syntagmatic relations (shown in italics in Table 2). This response pattern may in fact be tied to the gender biases observed, since syntagmatic relations are arguably more subjective than paradigmatic relations (related to context rather than logical relationships), and therefore leave more room for biased perceptions.

**Figure 2.** Concreteness effects for the original data (top), mistral data without profiles (center), and mistral data with profiles (bottom). Plots on the left display cue concreteness vs. number of associations, while plots on the right display cue concreteness vs. strength of associations.
**Concreteness effects.** The density plots in Figure 2 show cue concreteness vs. number of associations (left), and cue concreteness vs. strength of associations (right), for each of the three datasets. The expected concreteness effect [8] is present in the original data evidenced by the downward slope on the left (higher cue concreteness, fewer associates) and the upward slope on the right (higher cue concreteness, stronger associates). This effect appears to be absent in the Mistral-without-profiles data, and very subtle in the Mistral-with-profiles data.

**Network comparisons.** The networks shown in Figure 3 are subgraphs centered around the cue word *dog*, including only the top ten most frequent response nodes and the weighted directed edges from the cue *dog* to the responses. These visualizations demonstrate how the networks were built, and they also provide an idea of the types of differences that can be observed between the networks. For example, in the original network, the most frequent response to *dog* is *cat*, as evidenced by the very strongly weighted edge from *dog* to *cat*. Interestingly, *cat* is completely absent among the responses in the LLM subgraphs. Instead, the top responses are *bark* and *pet*. Another interesting property that we can observe is that the responses with paradigmatic relations to the cue appear to be the same in all three subgraphs (*puppy, canine, animal*(s), *pet*) while responses with syntagmatic relations to the cue seem to account for most of the variation among the responses (*cat, love, bone, furry, leash, loyal*).

![Network comparison](image)

**Figure 3.** Subgraphs centered around the cue word *dog* are show for all three networks. They include only the neighbors with the top ten highest in-degree (top ten most frequent responses). Only weighted directed edges from the cue *dog* to the responses are shown.

Table 3 shows the network statistics for the three networks constructed. We note that the LLM networks are sparser than the original networks. They also have much higher average edge weights and lower degrees compared to the original network. These statistics reflect the higher variation among responses in the original data compared to the LLM-generated data.

Table 4 shows the statistics that reflect the similarities and differences from the pairwise network comparisons. Regarding the pairwise node comparisons, there is a large percentage of node overlap in all three pairwise comparisons, evidenced by the high Jaccard coefficients. We observe that all nodes in the LLM networks are also in the original network, but that 7.8% and 19.5% of the nodes in the original network are not in the Mistral-without-profiles network and Mistral-
Statistics

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Mistral w/o profiles</th>
<th>Mistral w/ profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>nodes</td>
<td>12,228</td>
<td>11,271</td>
<td>9,841</td>
</tr>
<tr>
<td>edges</td>
<td>1,067,696</td>
<td>360,683</td>
<td>267,003</td>
</tr>
<tr>
<td>density</td>
<td>0.00714</td>
<td>0.00284</td>
<td>0.00276</td>
</tr>
<tr>
<td>&lt; k\textsubscript{in} &gt;</td>
<td>87.3</td>
<td>32</td>
<td>27.1</td>
</tr>
<tr>
<td>&lt; k\textsubscript{out} &gt;</td>
<td>241.9</td>
<td>259.7</td>
<td>244.8</td>
</tr>
<tr>
<td>&lt; w &gt;</td>
<td>2.77</td>
<td>8.11</td>
<td>9.02</td>
</tr>
<tr>
<td>&lt; C &gt;</td>
<td>0.078</td>
<td>0.092</td>
<td>0.079</td>
</tr>
<tr>
<td>&lt; L &gt;</td>
<td>2.72</td>
<td>3.89</td>
<td>4.47</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 3. Network statistics of the three networks are shown. The statistics include the numbers of nodes and edges, the network density, the average in-degree < k\textsubscript{in} > and out-degree < k\textsubscript{out} >, the average weighted in-degree < k\textsubscript{in}w > and weighted out-degree < k\textsubscript{out}w >, the average edge weight < w >, the average clustering coefficient < C >, the average shortest path length < L > and the network diameter < C >.

with-profiles network, respectively. These nodes represent cues that were never given as responses by the LLMs. Regarding the pairwise edge comparisons, we observe that there is very little overlap between sets of edges, especially between the original network and the LLM networks. The LLM networks are much more similar to each other than they are to the original network, with a Jaccard coefficient of 30% compared to 14.1% and 11.6%, respectively.

4. Conclusions and future work

We provided an overview of two novel LLM-generated word association datasets proposing some preliminary analyses that demonstrate how comparisons between human-generated and LLM-generated norms can be used to investigate various aspects of information processing. We find that human-generated responses are much richer and more varied than LLM ones. Also, Mistral-without-profiles responses are slightly more varied than Mistral-with-profiles ones, suggesting that more detailed prompts may limit response variability. We also observe stronger gender biases and weaker concreteness effects in the LLM-generated norms compared to the human-generated norms. In future work, we would like to expand our network analyses by exploring spreading activation processes on feature-rich networks to investigate the emergence of cognitive biases in humans and LLMs. Such investigations could have important implications for human-AI interaction.
References


Beyond Skills: The Role of Values in Job Seeking in the Era of Industry 5.0

Ajaya ADHIKARI\textsuperscript{a,1}, Quirine SMIT\textsuperscript{a} Joost A.G.M. VAN GENABEEK\textsuperscript{a}
Steven DHONDT\textsuperscript{a,c} and Cor J. VEENMAN\textsuperscript{a,b}

\textsuperscript{a}TNO, The Netherlands
\textsuperscript{b}Leiden University, The Netherlands
\textsuperscript{c}KU Leuven, Belgium

ORCiD ID: Ajaya Adhikari https://orcid.org/0000-0003-2518-7740, Quirine Smit https://orcid.org/0009-0008-8890-4261, Joost A.G.M. van Genabeek https://orcid.org/0000-0002-3873-7951, Steven Dhondt https://orcid.org/0000-0002-8888-5231, Cor J. Veenman https://orcid.org/0000-0002-3645-1198

Abstract. Jobseekers typically not only seek job vacancies matching their skills but also a company aligning with their values. This relates to Industry 5.0, a European Commission initiative emphasizing a more fulfilling role for workers. This study explores the relationship between skills and values in job vacancy selection and suggests several ways in combining these aspects for decision making. The first baseline system only uses skills, the second assigns equal importance to both skills and values, and the third, a hybrid intelligence system, leverages Pareto Optimality, leaving the ultimate decision on the trade-off between skills match versus values match to the jobseeker. Additionally, a small scale user study explores the impact of values on vacancy selection and evaluates the proposed matching systems. The results show that, participants seek a balanced trade-off between both skills and values. Accordingly, systems considering both skills and values outperform the baseline system. The system with equal weights and the Pareto optimality-based system have similar performances, possibly due to the large overlap in their output. Future work with more participants in a real-world application is needed to further validate our first exploration of the relationship between skills and values.

Keywords. skills and values-based matching, Industry 5.0, labor-market, multi-criteria decision making, pareto optimality

1. Introduction

Industry 5.0 is an European Commission initiative aimed at reshaping EU industrial policy [1]. Addressing concerns about the prior focus on technology and profit in the Industry 4.0 policy [2] being insufficient [3], Industry 5.0 emphasizes that the industry should be more human-centric, sustainable, and resilient [1]. In the human-centric aspect of Industry 5.0, ensuring the well-being and meaningful engagement of workers is crucial. This is also seen in the labor market from the jobseekers perspective, particularly post-
COVID-19, where the significance of passion for work is growing [4]. Jobseekers are increasingly seeking roles that are meaningful and fulfilling. Moreover, jobseekers increasingly place importance on having their values match those of their potential employers [5,6].

Wanberg et al. [7] highlight seven key factors that impact jobseekers’ decisions on which job vacancies to pursue. These factors include skills-related aspects, alongside one called “social capital,” which reflects the values of the jobseeker. Social capital represents the social context in which the jobseeker operates and offers values that resonate with the jobseeker. However, determining how important skills-related aspects are in comparison to values for effectively matching jobseekers with job vacancies is difficult [7]. Are skills-related factors more crucial, or are value-related factors equally important?

This study explores the interplay between skills and values in job vacancy selection and suggests potential matching systems based on skills and values. Three matching systems are proposed to provide relevant job vacancies for jobseekers. The first baseline system only considers skills for matching, the second system assigns predefined equal importance to both skills and values, and the third system employs a hybrid intelligence approach, where the decision on the importance between skills and values is left to the jobseeker using Pareto optimization. At last, this work describes a small user study with 20 participants to explore the role of values on the selection of job vacancies, while also assessing the performance of the matching systems compared to the baseline system.

This paper is structured as follows. Section 2 describes the related work. In section 3, we elaborate on the different matching systems and the setup of the user study. Section 4 reports the results. At last, sections 5 and 6 contain the discussion and the conclusion of our work.

2. Related Work

In the domain of recruitment, skills-based matching emphasizes aligning jobseekers’ skills with the skills requirements of available positions, diverging from traditional methods that predominantly consider academic qualifications and sector-specific experience. Various works has been conducted by labor market entities [8,9] and academic studies [10,11] to facilitate skills-focused matching. These works, however, largely overlook the incorporation of personal and organizational values in the matching process, highlighting an important area for improvement in recruitment approaches.

When matching on both skills and values the problem becomes a Multi Criterion Decision Making (MCDM) problem. Rather than optimizing for only one criterion (skills), the best solution (vacancy) is found by matching on two criteria (skills and values). We highlight three MCDM methods each with different advantages and disadvantages: self-learning, predefined weights and Pareto Optimization.

Self-learning AI methods are widely used in decision-making processes [12]. They can assist decision-makers or automate decisions entirely [13]. For our use case an AI model could be trained on past jobseeker-vacancy matches to identify characteristics that predict good matches. Its advantage lies in being able to automatically learn the relation between skills and values. It, however, assumes that future choices mirror past ones, which may not be applicable in our use case. If jobseekers’ priorities shift, the model may not adapt, relying on outdated preferences. It also assumes similar decision-
making among jobseekers with comparable profiles. At last, obtaining the necessary data, including vacancy details and jobseeker information such as application decisions and interview invitations, is difficult due to its sensitive nature and the volume needed.

The second method involves predefined weights, where users, in this case jobseekers, rate the relative importance of skills and values in advance. These ratings are then used to compute ideal matches [14]. This method is mathematically straightforward and can offer users insights into their priorities. However, obtaining accurate user weights can be challenging [14,15], as people may struggle to articulate their true preferences [7].

The third method, uses Pareto optimality [16] to select relevant job vacancies. Unlike other methods, Pareto optimality method does not assign weights to objectives. Instead, it offers users a variety of optimal solutions, enabling the users to choose without the need for predefined weights. While easily understandable in two dimensions, the method’s explainability diminishes with increased complexity in higher dimensions.

In this study, we include the method solely based on skills as a baseline and compare it to other two methods based on predefined weights and Pareto optimality.

3. Methodology

This section describes the methodology, implementation choices and user study setup for the pipeline that matches jobseekers with job vacancies, as illustrated in in Figure 1. The subsections elaborate on the various components of the pipeline.

3.1. Template of skills and values

This section explains how we create a template for skills and values. This template is used to describe both a job vacancy from an employer’s perspective and a jobseeker’s resume, as illustrated on the left side of Figure 1. For jobseekers, the template lists their skills and values. For job vacancies, it outlines the company values and the required job skills.

We use the European skills standard ESCO v1.1.12 to define a template of skills. ESCO defines a hierarchy of skills which are connected to different occupations. From this hierarchy, we select a subset of 24 skills at level 2. These skills fall uniformly under level 1 skills such that we get a broad selection of skills. These skills are specific enough for jobseekers to assess their own ability, yet not too specific that they don’t apply to many occupations. A jobseeker’s or job vacancy’s skills profile is represented by a vector of length 24 with possible numeric items of 0, 0.5, or 1, indicating the possessed or required skill level, respectively.

We define a template for values aligned with Industry 5.0, drawing from a recent report on the conceptual framework of Industry 5.0 [17]. Ten values are chosen and slightly adapted for our application such that jobseekers have a clear idea about the company values. To quantify the values profile for a jobseeker or job vacancy in a vector of size 10, we use a ranking ranging from 0.1 (least important) to 1 (most important), prompting individuals and companies to explicitly express the relative importance of their values. The list of the chosen skills and values are shown in the first column from the row 17 to

40 and 5 to 14, respectively, in Figure 2. A vector with 34 (24 skills and the 10 values) numeric items is regarded as a profile of a jobseeker or a job vacancy.

3.2. Generation of synthetic vacancy data and gathering of participant data

On the top left-hand side of Figure 1 the vacancy data is portrayed. For this study synthetic vacancies were generated to be suggested to the jobseekers based on ESCO. For each occupation in ESCO a synthetic job vacancy is created. The 24 skills defined in section 3.1 are considered as the skills profile of these job vacancies. Each skill gets a value of 1, 0.5 or 0, if the skill is present in the essential list, optional list or not present in the skills profile of the corresponding ESCO occupation. The 10 values are randomly ranked for each vacancy. This results in 3007 synthetic vacancies, each comprised of an ESCO occupation skills profile and a randomized company values profile. Figure 2 show two examples of job vacancies with their skills and values profiles in the columns C and D.

The other half of the matching system’s input, located on the lower left-hand side of Figure 1, comprises the jobseekers’ data. As a first exploration to validate the pipeline of Figure 1, we collect skills and values profiles from our colleagues within our data science research group. In line with ethical user-study guidelines of our institution, participants were informed about the research goals, the intended use of their data, and were asked for their voluntarily consent to participate. Thirty participants completed a questionnaire outlining their skills and values using the template in Section 3.1. The participants in this user study are primarily Data Scientists with at least a bachelor’s degree, all currently employed, and as far as we know, not actively seeking new employment. They were asked to imagine that they were looking for employment. For simplicity they are referred as “jobseekers” in this paper.

3.3. Matching systems

Given the skills and values profiles of a jobseeker and a list of vacancies, a matching system can select a subset of vacancies which are relevant to the jobseeker as shown in Figure 1.

For each jobseeker, two scores per vacancy are calculated; the skills-matching score and values-matching score. The skills-matching score is calculated by taking the cosine similarity between the skills vector of the jobseeker and the skills vector of a vacancy.
Figure 2. An example of skills and values profile of a jobseeker and two suggested vacancies in Excel format. The first column lists the different value names (row 5-14) and the different skills names (row 17-40). The second column shows self-reported skills and values profile of the jobseeker. The rest of the columns show the suggested vacancies by the different matching systems including their company name, values matching score, values profile, occupation name, skills profile and skills matching score (in the user-study the jobseekers could get up to 15 job vacancy suggestions). To evaluate the suggested job vacancies, the jobseekers filled in rows 42, 43 and 44 in which they had to pick 3-5 companies of which they liked the values, occupation and vacancy the most.

Previous research has shown the effectiveness of the cosine similarity in the context of calculating skills-matching scores [10]. The values-matching score is computed similarly by calculating the cosine similarity between the values vector of the jobseeker and the values vector of a vacancy. Given the two scores of each vacancy, we define three matching systems which select a sub-set of relevant vacancies of size $n$. Figure 3 visualizes an example of the three matching systems with $n = 10$.

The baseline, **skills-based** matching system only considers the skills-matching score and selects the top $n$ vacancies with the highest skills-matching score. This method does not consider the values matching score. Looking at Figure 3 (a), we see that the vacancies which are highest on the x-axis are selected.

The second, **predefined weights-based** matching system considers both skills and values-matching scores by assigning them predefined equal weights. A matching score is computed by taking the average of the two scores. The top $n$ vacancies with the highest

<table>
<thead>
<tr>
<th>Values and Skills</th>
<th>Job-seeker</th>
<th>Vacancy 1</th>
<th>Vacancy 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Company 1</td>
<td>Company 1</td>
<td>Company 2</td>
</tr>
<tr>
<td>Values matching score</td>
<td>0.39</td>
<td>0.15</td>
<td>0.40</td>
</tr>
<tr>
<td>provides training and education opportunities</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>values workers participation in decision making processes</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>values workers interests</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>values creativeness</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>values innovation</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>includes human-centric-values (such as diversity and inclusiveness) in business model and kpis</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>ensures human oversight in automated processes</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>makes and promotes ‘green choices’</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>designs and implements circular processes</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Occupation name</th>
<th>Component Engineer</th>
<th>Geographer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skills matching score</td>
<td>0.763</td>
<td>0.786</td>
</tr>
<tr>
<td>conducting studies, investigations and examinations</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>using digital tools for collaboration, content creation and problem solving</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>writing and composing</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>advising and consulting</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>providing information and support to the public and clients</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>organizing, planning and scheduling work and activities</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>developing objectives and strategies</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>programming computer systems</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>analyzing and evaluating information and data</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>raising and networking</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>raising and lifting</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>allocating and controlling resources</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>using digital tools to control machinery</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>using hand tools</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>documenting and recording information</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>using precision instrumentation and equipment</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>protecting and enforcing</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>providing health care or medical treatments</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>installing interior or exterior structures</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>installing interior or exterior infrastructure</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>installing machinery for the extraction and processing of raw materials</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>maintaining and repairing electrical, electronic and precision equipment</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>building and repairing structures</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

I like these companies’ values the most (pick 3-5) I like these roles/occupations the most (pick 3-5) These vacancies suit me the most (pick 3-5)
matching scores are selected. Looking at Figure 3 (b), we see that the vacancies in the top right corner are selected.

Third, we consider the multi-criteria system based on Pareto optimality. In this case, no prior weights between skills match and values match scores are assumed. The **Pareto optimality-based** matching system provides a sub-set of vacancies for which one dimension (skills-matching or values-matching score) cannot improve without the other dimension worsening. In Figure 3 (c), we see a spread of selected vacancies in a curved shape along the highest points on the y-axis and x-axis. The $p$ amount of Pareto optimal points, however, can be variable per jobseeker. To get a fixed amount of vacancies per jobseeker, we sample a random sub-set of size $n$ from the Pareto optimal vacancies if $p$ is greater than $n$. When $p$ is smaller then $n$, the Pareto optimal points are complemented with $n - p$ other vacancies. These other vacancies are randomly sampled from the vacancies that are Pareto optimal after removing the initial optimal vacancies. This step is repeated if the amount of the new Pareto optimal points is less then $n - p$. The Oapackage 2.7.13 python library [18] was used to compute the Pareto optimal points.

![Figure 3. An example of the three matching systems. It visualizes the 10 selected vacancies by each system from a large pool of synthetic vacancies for a jobseeker. The x-axis shows the skills-matching score and the y-axis the values-matching score. Note that the x-axis goes from 0 to 1, while the y-axis ranges from 0.55 to 1.](image)

3.4. Evaluation of Best Matches

In order to evaluate which of the three systems works best, the top 5 vacancies for all three of the methods were collected and presented to the jobseekers. Due to overlap of the selected vacancies per system, jobseekers saw 5 to 15 vacancies. The vacancies were presented to the jobseekers in an Excel sheet as the example shown in Figure 2.

Note that only the skills-based and the predefined weights-based matching systems provide an ordering of the results while the third does not. We discard the ordering and present the 5 to 15 vacancies in a random order. Jobseekers then objectively evaluate the suggested vacancies without being able to trace them back to the various matching systems.

Jobseekers were shown their self-reported values and skills and the vacancies suggested by all three matching systems. Each vacancy is made up of a company with values and an occupation with skills. For the values part of the vacancy a synthetic and anonymous company name was presented, as well as a values matching score and the score for each value. For the skills part of the vacancy the ESCO occupation name was shown as well as a skills-matching score and and its corresponding ESCO skills profile.
In the evaluation, the jobseekers answered three separate questions regarding which companies, occupations and vacancies they liked the most. The statements are: (1) *I like these companies’ values the most*, (2) *I like these roles/occupations the most*, and (3) *These vacancies suit me the most*. For each statement the jobseekers were asked to pick, respectively, 3 to 5 companies, occupations and vacancies which corresponded to the statement best. A total of 20 jobseekers completed this second part of the questionnaire.

### 4. Results

First, we examined the influence of values and skills on jobseekers’ choices of suitable vacancies in the user study. By assessing the correlation between the chosen companies’ values and the chosen vacancies, we gauged the impact of values on their choice of vacancies. Similarly, the correlation was determined for skills by comparing chosen occupations to vacancies. The Pearson correlation between chosen company values and vacancies was 0.2 ($P < 0.001$), while the correlation between chosen occupations and vacancies was 0.44 ($P < 0.001$). The chosen vacancies are twice as highly correlated with the chosen occupation skills than with the chosen company values. At the same time, we see that both chosen company values and chosen occupation have a low correlation with the chosen vacancies.

Second, we evaluated the performance of the three matching systems. Note that only the results from the three matching systems were shown to jobseekers, and they are required to choose a minimum of three vacancies per statement. This means that the evaluation reflects the relative performance of the systems in comparison to each other, rather than their general performance. Per jobseeker, we calculated the percentage of the suggested vacancies by each matching system that the jobseeker finally selected. On average, jobseekers chose 39.3% of the vacancies suggested by the skills-based system, 50.8% from the predefined weights-based system, and 49.2% from the Pareto optimality-based system. Note that the sum of the three percentages are higher than 100% due to the overlap between the suggested vacancies by the different systems. On average, there was a 0.75 out of 5 vacancies, or 15%, overlap in selected vacancies by the skills-based and predefined weights-based systems. There was a 1.45 out of 5 vacancies, or 29%, overlap between the skills-based and Pareto optimality-based systems. Lastly, there was a 1.95 out of 5 vacancies, or 39%, overlap between the predefined weights-based and Pareto optimality-based systems. The systems considering both skills and values outperform the skills-based system. The predefined weights-based system and the Pareto optimality-based system have similar performances. This might be due to the large overlap of 39% between the output of the both systems.

### 5. Discussion

This study shows that adding company values as additional selection criterion next to skills for job selection, in line with the Industry 5.0 initiative, fits the needs of jobseekers. Still, the current study has some limitations that need consideration.

First, the scale of the user study was limited, which impacts on the significance of the results. Second, our participants were employed data scientists, presumably not seek-
ing new jobs, which may influence their vacancy evaluations differently than active job seekers. Third, the users could only select from the ranked and selected vacancies by the proposed methods. The sheer volume of vacancies made it infeasible for a comprehensive evaluation of all listings. Accordingly, we could not evaluate whether or not there were vacancies outside the selected sets that would have been preferred by the participants. Finally, our study used synthetic vacancies. Real companies already have an image possibly supported by marketing appearance. Such an image may support, conflict or at least differ from their real company values. In this experiment the users were not influenced by a company image. However, the skill sets were accompanied by an occupation name. The image of these occupations could influence the opinions of the participants of the study.

Future research could explore if these results apply in real-world settings, on a larger scale, and with a broader audience. Moreover, in a real-world setting, there is a need to develop methods for accurately assessing the values of both companies and job seekers. This is difficult because there is often a discrepancy between what people and companies say they value, what they really value and what others perceive them to value. Furthermore, while our study addresses a bi-criterion problem, subsequent research could delve into more detailed analyses by assessing matches based on specific skills and values, either individually or in groups. Another research path could involve the enhancement of matching methods through interactive visualizations, allowing users to customize weightings based on personal preferences. Other similarity measures besides cosine similarity, like metric learning, can also be explored. Finally, our research primarily viewed the job-seeking process from the candidates’ perspective. Future studies could investigate the applicability of our methodology in aiding recruiters and companies in the selection process among numerous applicants.

6. Conclusion

This work investigates the relationship between skills and values in the process of selecting job vacancies and proposes possible systems for matching based on both skills and values. Three matching systems to match jobseekers with job vacancies are proposed: one considers only skills, another gives predefined equal weights to skills and values, and a third uses a hybrid intelligence approach based on Pareto Optimality, letting jobseekers decide the relative importance of skills and values. In addition a user-study was conducted to evaluate the relative performance between the matching systems and explore the relationship between values and skills when matching jobseekers with synthetic vacancies.

The results show that skills play a bigger role than values when choosing a vacancy. At the same time, it shows that the best job vacancies only based on skills is often not chosen. Participants seek a balanced compromise between matching both skills and values. Accordingly, systems that consider both skills and values outperform the baseline skills-only system. The system that assigns equal weights and the system based on Pareto optimality perform similarly, likely because there is a significant overlap in their output.
References


Formal Specification of Actual Trust in Multiagent Systems

Michael AKINTUNDE a,1, Vahid YAZDANPANAH b, Asieh SALEHI FATHABADI b, Corina CIRSTEA b, Mehdi DASTANI c and Luc MOREAU d

a King’s College London, London, United Kingdom
b University of Southampton, Southampton, United Kingdom
c Utrecht University, Utrecht, The Netherlands

ORCiD ID: Michael Akintunde https://orcid.org/0000-0002-5031-8813, Vahid Yazdanpanah https://orcid.org/0000-0002-4468-6193, Asieh Salehi Fathabadi https://orcid.org/0000-0002-0508-3066, Corina Cirstea https://orcid.org/0000-0003-3165-5678, Mehdi Dastani https://orcid.org/0000-0002-4641-4087, Luc Moreau https://orcid.org/0000-0002-3494-120X

Abstract. This research focuses on establishing trust in multiagent systems where human and AI agents collaborate. We propose a computational notion of actual trust, emphasising the modelling of an agent’s capacity to deliver tasks. Unlike reputation-based trust or performing a statistical analysis on past behaviour, our approach considers the specific setting in which agents interact. We integrate nondeterministic semantics for capturing inherent uncertainties within the behaviour of a multiagent system, but stress the importance of verifying an agent’s actual capabilities. We provide a conceptual analysis of actual trust’s characteristics and highlight relevant trust verification tools. By advancing the understanding and verification of trust in collaborative systems, this research contributes to responsible and trustworthy human-AI interactions, enhancing reliability in various domains.

Keywords. Trust, Multiagent Systems, Human-AI Interactions

1. Introduction

We are seeing a rapid adoption of AI agents being used for safety-critical tasks in the real-world and interacting with humans. A crucial step towards having symbiotic, responsible and trustworthy human-AI partnerships [1] is through the development of computational tools and methods to reason about how different components of such systems trust each other. We focus on actual trust 2, defined in terms of agents’ capacity to deliver tasks. Specifically, this research emphasises the significance of establishing trust in multiagent systems (MAS), where human and AI agents collaborate to achieve shared tasks. We propose a novel perspective on trust, wherein a trustee agent or group, referred to as β,
is considered trusted by another trustor agent or group, referred to as \( \alpha \), with respect to a specific task, denoted as \( T \), if \( \alpha \) can verify that \( \beta \) has the necessary strategic ability and epistemic capacity to successfully accomplish \( T \), and \( \beta \) has the intention to accomplish \( T \). This view on trust in MAS complements modelling trust solely based on an agent’s reputation or through a statistical analyses of historical behaviour. Instead, it emphasises the importance of considering the actual setting in which agents interact. Although statistical methods can be employed to narrow down the list of trusted agents for a given task, our approach underscores the significance of verifying a collective’s true capability to deliver in the current context.

Drawing inspiration from Halpern [3], we advocate for distinguishing history-based retrospective reasoning from prospective reasoning about what agents can actually ensure in a given setting, integrating formal logic-based methods within the framework of a MAS. In particular, we make a distinction between what is typically delivered by agents and what agents actually (i.e. they have the ability and intention to) deliver, and hence trusted for in a given setting (which can be verified based on the agents’ available actions and how such actions affect the system at hand and its properties). To that end, we argue that true trust verification necessitates an assessment of what agents are genuinely capable of accomplishing. Consider an autonomous delivery vehicle (ADV) tasked with transporting goods. Even if it was successful in former deliveries, it may currently have a low battery and is unable to achieve tasks, so the level of trust one has for the ADV needs adjustments based its current situation, regardless of its previous capabilities. We propose a computational notion of actual trust in MAS, which encompasses a comprehensive conceptual analysis of its key characteristics. We highlight the relevance of various formal verification tools that can be employed in MAS to ensure the delivery of trustworthy outcomes. We relax the strong assumption common in modelling MAS that agents have full observability; our trust model captures real-life uncertainties by assuming imperfect information.

This contributes to the ongoing exploration of trust in MAS by providing a robust computational foundation for assessing and verifying trust between human and AI agents. By adopting this perspective, we anticipate advancements in the development of responsible and trustworthy collaborative systems, paving the way for more effective and reliable human-AI interactions in various domains.

2. Related Work

We remark that while the work of [4,5,6,7] assumes the existence of trust relations at design-time, and allows for reasoning about more complex trust dynamics on top of the trust relations they assume, we formulate trust as a notion which emerges through the dynamic evolutions of the multi-agent system. Their trust relation is static but we allow modelling and reasoning about trust in given worlds (global states) of the MAS. Differently also to their work, we consider intentions, although the set of intentions that we consider for our trust semantics is static. Following Cohen and Levesque [8], we refer to the set of intentions as goals that an agent or agent group has chosen and is committed to delivering. In the rest of the text, we may refer to the elements of this set as intended goals or simply as goals. In comparison to reputation-based methods with a retrospective approach to trust [9,10], we maintain a prospective view of trust and build
trust based on agents’ ability, their knowledge of the environment and what they intend to achieve in a MAS. Different to our focus on trust, a complementary line of work is the development of methods to formalise notions of responsibility in groups such as in [11], where strategic, probabilistic and temporal modalities are used to reason about a group’s responsibility for taking risks. We also remark in comparison to the more abstract logic of [12], our focus here is to ground the logic on a suitable computational model and to be able to apply automated verification techniques to reason about trust in a computationally feasible manner. We do not consider beliefs in the sense of [13] in our modelling of trust.

2.1. Conceptual Analysis: Trust as a Prospective Concept

Trust between groups of agents is inherently a social phenomenon, with multiple defining characteristics such an agent’s ability and knowledge, which also exhibits a temporal dimension [10]. Trust in MAS has been widely studied in the literature, which we find centers on three dominant themes:

Cognitive Trust Modelling. One influential perspective on trust is presented in [14], where trust is modelled based on the cognitive states of agents. Through this, an agent can be trusted if their beliefs and intentions, and accordingly their plan of action is aligned with our intended plans and intended future. However, in complex systems like human-AI interactions, determining an agent’s true intentions can be challenging, and intention elicitation remains an open problem in AI systems. While cognitive models of trust offer a high-level understanding of trust dynamics, their implementation in large-scale human-AI settings, such as smart mobility applications, requires further research.

Reputation-Based Trust. Another approach to trust, proposed by [15], focuses on past behaviour and agents’ reputation. It suggests that agents who have consistently performed according to plans in the past can be trusted to deliver similar performance in the future. This is particularly applicable in domains with predictable environments, such as closed world databases where agents follow safe protocols for data updates. However, when considering trust for a specific task, it is more reliable to evaluate an agent’s available actions and knowledge in the current state of the MAS. What agents delivered in the past can be used as a means to limit the search space for trusted agents to deliver a particular task but may result in biased evaluations if used as the sole measure for establishing trust.

From Trust Relations to Collective Trust. More recent lines of research on trust [5] assumes trust relations and builds a framework that analyses how these assumed trust relations cascade in the system, leading to the formation of collective-level trust among coalitions. However, the core assumption of this framework is access to a social network of bidirectional trust relations. Although in some settings such a social network may be available and fixed, the assumption of fixed trust relations as given may limit the applicability of this approach in dynamic environments. More research in this line shows how different types of knowledge [16] and uncertainties regarding the epistemic state of agents and agent groups can be integrated into modelling trust dynamics.

Highlighted by [17] is the phenomenon of humans tending to place unwarranted levels of trust in AI systems during their interactions. This has been observed in studies of humans interacting with robots in high-risk scenarios [18]; humans can “overtrust”
robots when they are observed to malfunction during prior interactions. This underscores the need for the development of methods aimed at verifying the trustworthiness of a specific AI agent within a particular context, rather than making generalised assumptions based solely on past interactions. This holds significant promise in refining the trustworthiness of AI systems and mitigating instances of unjustified trust solely rooted in historical interactions.

Against this background, we believe it is crucial to distinguish between two types of trust: retrospective trust that reasons about trusting an agent based on past and prospective trust which looks at the abilities of agents and what they can deliver in the future. We denote the former as typical trust (as it relies on the typical, historical, and statistical data on agent’s past performance and reputation) and the latter as actual trust as it relies on the actual state of the system and what the agent can deliver under their strategic and epistemic limitations. We ideate that trusting an agent solely based on historical data disregards the contextual factors and the specific requirements of the current scenario. Instead, a more comprehensive approach to establishing trust is needed, which is not solely based on the performance of an agent’s past behaviour but also relies on an agent’s ability to deliver a task in the future, determined by the current system state. This can be modelled as a concurrent game, where trust in a group of agents with respect to a specific task can be determined based on the foreseeable consequence of joint actions. An intelligent agent can utilise both actual and retrospective trust notions to obtain a comprehensive evaluation of who to trust.

As a motivating example, The Bit Transmission Problem [19] underscores the limitations associated with relying solely on an agent’s past behaviour to establish trust. This problem involves the transmission of the value of a bit from a (potentially human) source agent to a destination agent. Traditional trust models, which hinge on reputation-assessments and past behaviour, would suggest that an agent with a proven track record of successfully communicating information can be considered trusted for future transmissions. However, this approach fails to consider the dynamic nature of the system and the specific contextual factors at play. It is crucial to evaluate the agents’ current capacities (as actions they possess and how their actions may affect the environment), knowledge (as what they know about their own ability and can distinguish in possible next states of the system), intentions of the agents, and the specific requirements of the task within the existing system state. By examining the agents’ available actions and knowledge within the present system configuration, a more informed judgement can be made regarding their reliability in transmitting bits. Therefore, a shift in focus is necessary, moving away from relying solely on historical performance metrics and towards verifying agents’ present capabilities and potential to deliver desired outcomes, facilitating a comprehensive verification of trust.

The exploration of trust verification techniques has higher-level consequences that extend beyond individual assessments of trust in MAS. While reputation-based trust models provide valuable insights, it is important to acknowledge that relying solely on reputation may result in biased misplaced trust. By delving into the intricacies of trust and developing reliable verification methods that look at temporal dynamics of trust and allow reasoning about what agents can deliver prospectively, we can pave the way for ethical AI systems. These systems allow users to foresee what other (AI) agents can de-
liver, compare those against their own preferences and values, and actively reason about and verify trust in a reliable and transparent manner. To address these critical aspects, we require temporally expressive tools to represent and reason about trust dynamics. In the next section, we survey such formal tools and highlight weaknesses and points we learned for formulating our perspective on actual trust in MAS.

3. Formal Methods for Temporal Trust Reasoning

We re-iterate that actual trust is forward-looking and requires formal methods for temporal reasoning about future computations occurring in a MAS. We therefore direct our focus here on works where temporality is a key aspect of reasoning about trust. To this end, little is explored in the literature for performant techniques for verifying temporal trust properties in MAS. Drawel et al. [4] address the problem of model checking trust logics, and propose a model checking technique for two temporal logics for trust, namely TCTL, a (pre-conditional) Computation Tree Logic of Trust and TCTL$^{c}$ for conditional trust, which extend CTL to introduce new modalities to reason about trust. They ground the semantics of their trust logics on a formalism of interpreted systems [19] which is enriched with a trust function – a binary relation between two states, which associates to each local state of each agent the “trust vision” the trustor has towards other agents in a given global state. States are compatible with each other with regard to the trust an agent has with another. The (pre-conditional) trust modality $T_p(i, j, \psi, \varphi)$ in TCTL stands for “Preconditional Trust” and is read as “the trustor $i$ trusts the trustee $j$ to bring about $\varphi$ given that the precondition $\psi$ holds”; $\psi$ holds in a state $s$, and $\varphi$ needs to hold in all trust-accessible states between agents $i$ and $j$ different from $s$. On the other hand, the conditional trust modality $T_c(i, j, \psi, \varphi)$ in TCTL$^{c}$ is read as “agent $i$ trusts agent $j$ about the consequent $\varphi$ when the antecedent $\psi$ holds”. Differently from pre-conditional trust where the precondition $\psi$ must hold before the trust content $\varphi$ is brought about, conditional trust needs at least one trust-accessible state satisfying $\psi$. Agents are specified in their model using the VISPL [6] language which extends ISPL [20] used for specifying traditional interpreted systems. A transformation-based procedure is used to convert the model checking problem of TCTL and TCTL$^{c}$ into model checking CTL. New symbolic model checking algorithms are proposed to formally specify and automatically verify a system under consideration against properties in TCTL and TCTL$^{c}$. The semantics of their trust logics are interpreted in a vector-extended version of interpreted systems, capturing the trust relationships between interacting parties within the model itself. We argue that assuming trust relations as known in advance is not realistic and is a phenomenon to be reasoned about automatically as we do here. For model checking, transformations to CTL-supported models and formulae are exploited to make use of the NuSMV model checker CTL [21].

Inspired by Falcone and Castelfranchi’s [14] cognitive notion of social trust, Huang and Kwiatkowska [13] introduce a framework for quantifying and reasoning about cognitive trust, governing social relationships between humans and autonomous systems. A semantic model grounded on stochastic games is introduced, namely Autonomous Stochastic Multi-Agent Systems (ASMAS). Differently to [4] where trust is represented as a binary relation, probabilistically quantified degrees of trust are expressed in terms of belief. The ASMAS model combines a stochastic game, where agents only have partial
observability of the state, with a mechanism providing agents with goals, intentions and preferences. An ASMAS differentiates between transitions in the temporal and cognitive space; the actions of the agent occur in physical space, whereas cognitive processes lead to changes in the agent’s mental state, and lead to decisions about which (physical) action to take. Properties are specified using PRTL* (Probabilistic Rational Temporal Logic), an extension of the probabilistic temporal logic PCTL* which introduces cognitive attitude and trust operators. In PRTL*, cognitive operators $\mathcal{G}_A \psi$ ($\psi$ holds in the future regardless of agent $A$ changing its goals), $\mathcal{I}_A \psi$ ($\psi$ holds in the future regardless of $A$ changing its intentions), $\mathcal{C}_A \psi$ (agent $A$ can change its intention to achieve $\psi$), are used with respect to an agent $A$ and a task $\psi$; they quantify over the possible changes of goals, intentions and available intentions respectively. For $\bowtie \in \{<,\leq,>,\geq\}$ and $q \in \{0,1\}$, the belief operator $\mathcal{B}_A^{\bowtie q} \psi$ (agent $A$ believes $\psi$ with probability in relation $\bowtie$ to $q$) probabilistically quantifies agent $A$’s belief of $\psi$, the competence trust operator $\mathcal{CT}_A^{\bowtie q} \psi$ (agent $A$ trusts agent $B$ with probability in relation $\bowtie$ with $q$ on its capability of completing task $\psi$) probabilistically quantifies the degree of agent $A$’s trust in $B$’s capability to achieve $\psi$ and that there exists a valid intention for $B$ to carry out $\varphi$, and the disposition trust operator $\mathcal{DT}_A^{\bowtie q} \psi$ expresses that agent $A$ trusts agent $B$ with probability in relation $\bowtie$ on its willingness of completing the task $\psi$; that is, $\psi$ is unavoidable for all inten-
tions. These operators probabilistically quantify an agent’s beliefs, the degree of agent $A$’s trust in $B$’s capability and willingness to carry out $\psi$ respectively for each case.

As seen from the above works, verifiable notions of actual trust need to be forward-looking and require practical formal methods for temporal reasoning. Our learnings point us towards our formulation on actual trust in MAS.

4. Modelling Interactions in Multiagent Systems

In this section we discuss the basic building blocks necessary to reason about trust in a MAS. An interpreted system (IS) [19] is a model for representing a MAS; it is a formal description of the computations carried out by a set of agents.

**Definition 1 (Interpreted Systems).** For a set of agents $Agt = \{1,\ldots,n\}$, where $n$ is the number of agents in the system, an interpreted system IS is a tuple: $IS = \langle (L_i, \text{Act}_i, P_i, t_i)_{i \in Agt \cup \{E\}}, I, h \rangle$, where:

- Each agent $i \in Agt \cup \{E\}$ is characterised by a finite set of private local states $L_i$, which determines all information relevant to agent $i$ in a given global state.
- $\text{Act}_i$ is a finite set of actions that may be performed for agent $i$.
- $P_i: L_i \rightarrow 2^{\text{Act}_i \setminus \emptyset}$ is a protocol for agent $i$. The actions of the agent must be performed in compliance with the protocol, which allows for non-determinism in the system when more than one action is enabled for a given local state.
- $E$ is a special “agent”, referred to as the environment. It has its own sets of local states $L_E$, actions $\text{Act}_E$, protocol $P_E$ and transition function $t_E$. 

• $t_i : L_i \times \text{Act}_1 \times \cdots \times \text{Act}_n \times \text{Act}_E \rightarrow L_i$ is a (partial) transition function describing the (deterministic) evolution of agent $i$'s local states.\footnote{Here we adopt the definition with $n + 1$ evolution functions as used in [22]. Each evolution function gives the next local state as a function of the current local state of the agent and all the other agents’ actions.} Every action is assumed to be protocol-compliant: if $l_i' = t_i(l_i, a_1, \ldots, a_n, a_E)$, then $a_i \in P(l_i)$ for all $i \in \text{Agt} \cup \{E\}$.

• $I$ is a set of initial global states.

• $h : \text{AP} \rightarrow 2^0$ is a valuation function defining the set of states where certain atomic propositions are true, where $\text{AP}$ is a set of atomic propositions and $Q = L_1 \times \cdots \times L_n \times L_E$ is a set of global states.

Local states cannot be observed by other agents. Actions are observable by other agents i.e. to determine an agent’s transition function $t_i$. Global states $Q$ combine the local states of all $n + 1$ agents. Global actions $\text{ACT} = \text{Act}_1 \times \cdots \times \text{Act}_n \times \text{Act}_E$ combine all agents’ action sets. Consider an agent $i \in \text{Agt} \cup \{E\}$. For a global state $q = (l_1, \ldots, l_n, l_E) \in Q$ the function $\text{loc}_i : Q \rightarrow L_i$ where $\text{loc}_i(q) = l_i$ returns the local state of agent $i$ in global state $q$, and for a joint action $a = (a_1, \ldots, a_n, a_E) \in \text{ACT}$ the function $\text{act}_i : \text{ACT} \rightarrow \text{Act}_i$ is such that $\text{act}_i(a) = a_i$ returns the action of agent $i$ in joint action $a$. A global transition function defines the means through which the system evolves through time.

**Definition 2** (Global Transition). Given two global states $g, g' \in Q$ and a joint action $a \in \text{ACT}$, a (partial) global transition function $t$ is such that $t(g, a) = g'$ iff $t_i(\text{loc}_i(g), \text{act}_i(a)) = \text{loc}_i(g')$ and $\text{act}_i(a) \in P(\text{loc}_i(g))$ for all $i \in \text{Agt} \cup \{E\}$. This property can be abbreviated $g \xrightarrow{a} g'$ iff $t(g, a) = g'$.

Since each $t_i$ is deterministic, at most one such transition exists for each global state $g$ and $a$. A joint action $a \in \text{ACT}$ is enabled in state $q \in Q$ if there exists a state $q' \in Q$ such that a transition exists between $q$ and $q'$ through the execution of $a$. A Kripke model $\mathcal{M}_{IS}$ associated to an interpreted system $\mathcal{IS}$ defines its semantics when interpreting temporal formulae. Knowledge of agent $i$ is represented in terms of an indistinguishability relation $\sim_i$, an equivalence relation with known properties: reflexivity, symmetry and transitivity. Informally, a property $\varphi$ being “known” by an agent $i$ is determined by $\varphi$ holding in all global states indistinguishable by $i$.

**Definition 3** (Associated Kripke model). A Kripke model $\mathcal{M}_{IS}$ associated with IS = $(\{l_i, \text{Act}_i, P_i, t_i\}_{i \in \text{Agt} \cup \{E\}}, I, h)$ is a tuple $\mathcal{M}_{IS} = (W, R, h, \{\sim_i\}_{i \in \text{Agt}})$ such that worlds $w \in W$ are the global states of IS reachable via the set of initial states $I$ through the transition relation $R$. Two worlds $w, w' \in W$ are related by the transition relation $R \subseteq W \times W$ when there is a joint action $a$ such that $w \xrightarrow{a} w'$. The valuation function $h$ is used as a labelling function, and the epistemic indistinguishability relation [19, p. 117] for agent $i \in \text{Agt}$, namely $\sim_i \subseteq W \times W$, relates a pair of global states $w, w' \in W$ whenever agent $i$ has the same local state in both $w$ and $w'$. In other words, it cannot distinguish $w$ and $w'$. A (potentially infinite) path $\pi$ is defined as a sequence of states where each pair of successor states are related by a global transition; that is, a path is a sequence of states $\pi = (q^0, q^1, \ldots, q^r, \ldots)$ such that for all $i \geq 0$, we have $(q^i, q^i, q^{i+1}) \in t$ for some action $a_i', \ldots, a_E' \in \text{ACT}$. The state at position $k$ is denoted $\pi(k)$. $Q_T$ denotes the projection of $Q$ on the local states of the agents in $\Gamma \subseteq \text{Agt} \cup \{E\}$ and similarly $\text{ACT}_T$ for the elements of $\text{ACT}$ restricted to the agents in $\Gamma$. For example, if $a = (a_1, a_2, a_3, a_4, a_5, a_E) \in \text{ACT}$ and we
take $\Gamma = \{1, 3, 5\}$, then an element $a_\Gamma$ of $\text{ACT}_\Gamma$ is $a_\Gamma = (a_1, a_3, a_5)$. A strategy provides the semantics in an IS of strategic operators in temporal formulae such as those encountered in ATL [23].

**Definition 4 (Strategy).** For an agent $i$ of an IS, a (memoryless) strategy $s_i$ is a function $s_i : L_i \to 2^{\text{ACT}_i} \setminus \{\emptyset\}$ such that if $a_i \in s_i(l_i)$, then $a_i \in P_i(l_i)$.

A strategy depending on a history, or sequence, of local states is known as a memory-based strategy. In addition to this, note that agents may perform different actions in different global states whose local component is the same, allowing for non-deterministic strategies. This is known as a non-uniform strategy. As is done in [20], we also focus on non-uniform, memoryless, incomplete information strategies here. In a MAS, joint strategies are a collection of individual strategies.

**Definition 5 (Joint Strategy).** Given a coalition $\Gamma$, a joint strategy for $\Gamma$ is a function $s_\Gamma : Q_\Gamma \to 2^{\text{ACT}_\Gamma} \setminus \{\emptyset\}$ such that $s_\Gamma(l_{x_1}, \ldots, l_{x_k}) = (s_{x_1}(l_{x_1}), \ldots, s_{x_k}(l_{x_k}))$, where $s_{x_1}, \ldots, s_{x_k}$ are strategies for the agents $x_1, \ldots, x_k \in \Gamma$.

**Intentions in MAS.** We consider intentions as goals that each agent intends to deliver. They are declared publicly in terms of statements in propositional logic.

**Definition 6 (Interpreted systems with intentions).** We define an interpreted system with intentions as a tuple $\text{ISI} = (\langle (L_i, \text{ACT}_i, P_i, t_i, I_i) \rangle_{i \in \text{Agt} \cup \{E\}}, I, h)$, where a consistent set of intentions $I_i \subseteq 2^\Phi \setminus \emptyset$ are such that each agent $i$ is associated to a finite set of $k$ propositions $\{\phi_1, \ldots, \phi_k\}$, with each $\phi \in \Phi$ being propositional formulae, that it intends to bring about irrespective of the global state of the system and irrespective of all strategies of any agent in the system.

We assume a consistency constraint on the set of intentions for individual agents intentions; $p \in I_i \Rightarrow \neg p \notin I_i$ for all $i \in \text{Agt}$, for any proposition $p$. It is not possible for an agent to have the intention to go out and stay at home simultaneously. The model $\mathcal{M}_{\text{ISI}}$ associated with the interpreted system with intentions $\text{ISI}$ is defined identically to that in Definition 3. Note that an interpreted system is a special case of an ISI, where $I_i = \top$ for all $i \in \text{Agt}$. Unlike in the treatment introduced in [24], intentions are not bound to states or strategies; intending to bring about one or more propositions is orthogonal to the agent’s ability to do so.

**Alternating-time temporal logic.** Alternating-time temporal logic (ATL) [23] generalises CTL and is used for strategic reasoning in MAS. It is used to describe what a collection of agents can achieve. Similarly to CTL, we have the usual atomic propositions, negation, binary conjunction and disjunction operators $p, \neg, \wedge$ and $\vee$ respectively. In ATL, given a set of agents $\text{Agt}$, a coalition $\Gamma \subseteq \text{Agt}$ and a property $\phi$, the specification $\langle \Gamma \rangle X \phi$ is read as: “the coalition $\Gamma$ have a joint strategy to achieve $\phi$ in the next step independently of what $\text{Agt} \setminus \Gamma$ does.”

Here and in [20] we interpret the semantics of ATL on the temporal model $\mathcal{M}_{\text{IS}}$, given an initial state $q^0 \in Q$, a formula $\phi$ and a set of atomic propositions $A_P$ where $p \in A_P$. The semantics for non-temporal operators are equivalent to that for CTL. We

---

5 When in the context of speaking about a collection of agents, we refer to joint strategies simply as strategies.
here will focus on the semantics of the “next” operator $X$ only: $(\mathcal{M}_{IS}, q^0) \models (\Gamma)X\varphi$ iff there exists a joint strategy $s^1$ and joint action $a^1 \in s^1(\mathcal{M}_{IS})$ such that for all actions $a$ whose restriction to $\Gamma$ is equal to $a^1$ and for all states $q^1$ such that $q^0 \rightarrow_a q^1$ we have that $(\mathcal{M}_{IS}, q^1) \models \varphi$. Connectives for reasoning about strategic ability over sequences of states are not discussed here.

ATLK [20] combines ATL with modal operators to reason about the knowledge of agents in a MAS. Here we focus on a fragment which is critical to define trust modalities consisting of those given in “Vanilla ATL” with the knowledge operator $K_i \varphi$, which is read as “agent $i$ knows $\varphi$”. More precisely, for a model $\mathcal{M}$, state $q^0$ and property $\varphi$, $(\mathcal{M}, q^0) \models K_i \varphi$ iff for all $q^1 \in Q$ we have that $q^0 \sim_i q^1$ implies that $(\mathcal{M}, q^1) \models \varphi$. In other words, for agent $i$ there isn’t a state indistinguishable from $q^0$ where $\varphi$ does not hold. The agent has enough information in its own local states to determine from its perspective that $\varphi$ holds in the system.

We note that we are modelling trust under perfect information. That is, what a group intends to do is known among the group members, so due to the public declaration of intentions, what a group intends to deliver is in a sense also what every individual within the group intends to do as well. Although we do not focus on complexity in this work, the semantics of ATLK in this context are analogous to the imperfect information, memoryless strategies case of ATL, namely ATL$_{no}$ [25,26]. The model checking problem for ATL$_{no}$ is decidable and $\Delta^2_P$-complete.

5. A Computational Notion of Actual Trust

The semantics defined in the previous section form the basis of that which will be used to reason about actual trust, where an agent trusts other agents to collectively perform a task, which we encode here in terms of ATLK formulae. This will aid in being able to transform from the trust verification problem into an ATLK model-checking problem. As common in formal methods research, we evaluate our notion’s properties formally and show applicability in a well-established running example to showcase the expressiveness of our formal notion of actual trust for reasoning about different aspects of trust in multiagent systems. We call the specification language $L$. Assume that $L$ contains the standard Boolean connectives of CTL. In terms of the trustee $\beta$ (a group of potentially trusted agents), the trustor $\alpha$ and task $T$ (see Section 1), we take an agent $i$ as $\alpha$, the group of agents $\Gamma$ as $\beta$ and our task $T$ as the formula $\varphi$. We assume the trust operator $T_i$ which takes as input an agent $i$, a group of agents $\Gamma$ and an $L$ formula $\varphi$. The formula $T_i(\Gamma, \varphi)$ is read as “agent $i$ trusts $\Gamma$ to bring about $\varphi$”. Specifically:

**Definition 7** (I trust $\Gamma$ if I know they can deliver). Given a model associated with an interpreted system with intentions $\mathcal{M}_{IS}$ and an agent $i \in \text{Agt}$, we say that $(\mathcal{M}_{IS}, q^0) \models T_i(\Gamma, \varphi)$ iff for all $q^K \in Q$ we have that if $q^0 \sim_i q^K$ then there exists a (collective) strategy $s^1$ for $\Gamma$, and action $a^1 \in s^1(\mathcal{M}_{IS})$ such that for all states $q^1$ such that $q^K \rightarrow_a q^1$, we have that $\varphi \land \bigwedge_{i \in \Gamma} I_i$ is nonempty and consistent, and $(\mathcal{M}_{IS}, q^1) \models \varphi$.

That is, $\varphi$ is consistent with each agent’s intentions. We note that with this definition, trust is defined in terms of what agents intend to deliver regardless of their ability to deliver; one may intend $\varphi$ regardless of its ability to deliver it from any local state. The intersection $\bigcap_{i \in \Gamma} I_i$ finds a consistent set of intentions that all agents intend to deliver.
It is permitted for $i \in I$ or $I = \{i\}$, where agent $i$ trusts that it can cooperate with the agents in $I$ to bring about $\varphi$, and that agent $i$ has trust in itself that it can bring about $\varphi$ respectively, regardless of what the agents in $Agent \setminus I$ do. From this, it is possible to reason about supersets of agents:

**Proposition 1** (Non-monotonicity of trust). Let $I \subseteq I'$. Then $T_{I'}(I', \varphi) \not\subseteq T_{I}(I, \varphi)$.

**Proof.** Assume a $\mathcal{M}_{ISI}$ with $Agent = \{1, 2, 3\}$, $I = \{1, 2\}$, and intentions $I_1 = \{\varphi, \psi\}$, $I_2 = \{\varphi\}$ and $\mathcal{I}_3 = \{\neg\varphi\}$. Each $I_i$ is clearly consistent. Now, without loss of generality, let $i = 1$ and assume $q = T_I(I, \varphi)$ for some state $q \in Q$. Assume $I' = I \cup \{3\} = \{1, 2, 3\}$. For $q = T_I(I', \varphi)$ to hold, it would mean that $\varphi \cap \bigcap_{i \in I'} I_i = \varphi \cap \{\varphi, \psi\} \cap \{\varphi\} \cap \{\neg\varphi\} = \emptyset$ is consistent, which is a contradiction.

This highlights the importance of considering intentions – without them, a notion of trust could show unintuitive results, but with the inclusion of intentions, not all supersets of a set of agents can be trusted for $\varphi$. We illustrate our notion of trust in The Bit Transmission Problem (BTP) [19, p. 114].

**Example 1** (Bit Transmission Problem). *In the bit transmission problem, a sender $S$ wants to communicate the value of a bit to a receiver $R$ over a faulty communication channel. Messages between $S$ and $R$ may be lost, but the value of the bit will not be corrupted. One needs to define a protocol for $S$ to be sure that $R$ received the bit. An example protocol is the following: $S$ sends the value of the bit to $R$ and continues to do so until it receives an acknowledgement (“ack”), after which it will stop sending the value of the bit. $R$ does nothing until it receives the value of the bit, and then it sends acknowledgements to $S$ forever in the future.*

We encode the BTP as an ISI in a standard way; by considering the local states of the sender and receiver agents $S$ and $R$, and the environment agent $E$, which will be used to represent the faulty communication channel. The sender’s state will consist solely of the value of the bit, or the value of the bit combined with the acknowledgement sent from the receiver. The local states representing the situations where the sender has received the acknowledgement are denoted $0\text{-}ack$ and $1\text{-}ack$. The receiver was either sent the value of the bit, with its local state equal to the bit’s value, or it is empty, denoted by $\varepsilon$. The environment state does not play a role in our formalisation of the bit transmission problem, so we take $L_E = \{\varepsilon\}$. We now have the following local states: $L_S = \{0, 1, 0\text{-}ack, 1\text{-}ack\}$, $L_R = \{0, 1, \varepsilon\}$ and $L_E = \{\varepsilon\}$. Omitting the environment’s local state, this generates six global states: $(0, \varepsilon), (0, 0), (0\text{-}ack, 0), (1, \varepsilon), (1, 1),$ and $(1\text{-}ack, 1)$. Consider the propositional atom $\text{recack}$, representing all global states where the receiver was successfully transmitted the bit’s value and the sender has received the acknowledgement, such that $h(\text{recack}) = \{(0\text{-}ack, 0), (1\text{-}ack, 1)\}$. Then, the interpreted system $IS$ consisting of the agents $S$, $R$ and $E$ satisfies $\text{recack}$ at any global state $q \in I$, where $I$ is the set of initial states containing the two global states with the sender’s local component of $q$ being either $1\text{-}ack$ or $0\text{-}ack$.

Assume in the corresponding ISI that all agents intend for acknowledgements to always be received, i.e. $I_S = I_R = I_E = \text{recack}$. It is easy to check whether $IS, q = T_S(R, \text{recack})$, i.e. the sender trusts the receiver in bringing about recack. Intuitively, starting from the initial state $q$ of either $(0, \varepsilon)$ or $(1, \varepsilon)$, the only states accessible via the sender’s accessibility relation $\sim_S$ are $(0, 0)$ and $(1, 1)$ respectively. For both of these...
states, there exists a strategy for $R$ such that regardless of what $S$ does, states satisfying \textsc{recack} follow, since the receiver can always send the acknowledgement to the sender regardless of the value of the bit in the previous state. If the faulty communication channel is modelled in the environment, and in every state it prevents the receiver from sending the acknowledgement, the formula would not hold, as a suitable strategy for $R$ does not exist, even though the environment intended that acknowledgements should always be received since $I_E = \textsc{recack}$ in the model.

Expanding upon this notion of trust, our modelling also facilitates reasoning about trusting agents who employ multistep strategies to ensure the realisation of a state of affairs $\varphi$. While our formalisation emphasises the immediate outcomes, our model acknowledges the potential for agents with longer-term strategies to be trusted as well. By verifying the effectiveness of multistep strategies, agents can evaluate and place trust in individuals or groups who demonstrate the ability to “eventually” achieve a desired outcome $\varphi$. This extended perspective on trust enables a more comprehensive analysis of trust dynamics in complex scenarios, accommodating both immediate and long-term strategies for attaining desired goals as well as providing a base for quantifying trust (e.g., agent $i$ may trust a $\Gamma$ who can ensure $\varphi$ in the immediate next state more than $\Gamma'$ with a multistep strategy to do so).

6. Discussion: Expressivity for Modelling Trust Dynamics

\textit{Trust is Bounded by Knowledge.} Actual trust is limited by an agent’s knowledge; an agent’s trust in other agents is dependent on the information it possesses and its ability to discern and evaluate the ability of others. We account for the relationship among states that an agent may not be able to differentiate due to its limited knowledge. For $T_i(\Gamma, \varphi)$ to hold, the trustee must have sufficient information to assess the potential consequences of the trusted agents’ actions and anticipate the states they will reach as a result. The trustor(s) must possess the necessary knowledge for the fulfilment of a task. We capture the epistemic dynamics of trust and applicability for reasoning about trust in real-world scenarios.\(^6\) We use the standard knowledge and strategic operators $K_a\varphi$ and $\langle a \rangle \varphi$ assuming the “de dicto” semantics of knowledge: an agent only knows that a strategy is available. Outside the scope of this work is to consider a stronger view of strategic ability, e.g. such as that introduced in [27] where an agent also knows the specific strategy.

\textit{Trusting Coalitions.} The relationship between individual- and collective-level trust is rooted in ATL and the semantic machinery that we used to model trust as it allows us to reason about collective-level capacities, knowledge of agent groups, and accordingly our notion of actual trust in MAS. Our notion is expressive enough to evaluate if for an agent $i$ trusting agent $j$ regarding a task $T$, whether it is reasonable to also trust any group $J$ including $j$ for delivering $T$. This requires considering whether their intentions are aligned on top of their strategic ability to deliver the task in question. Trust in an individual may not necessarily extend to encompass trust in larger groups including that individual. Our notion of trust allows for reasoning about the expansion of trust beyond the individual level, enabling us to consider trust dynamics within collective entities. The framework

\(^6\)We highlight that as we modelled our notions in ATL, verifying actual trust can be implemented in standard model-checking tools such as MCMAS [20].
of interpreted systems also allow for a group of external observers to be modelled either as the environment “agent” \( E \) in a similar sense to [28, p. 10] using the trivial protocol function (returning a no-op action for all local states). By recognising such relationships between individual and collective trust, we gain a better understanding of trust dynamics in human-AI systems. We can analyse how individual-level trust influences coordination within groups. This understanding is crucial in various contexts, such as teamwork, organisational dynamics, and social networks, where trust plays a pivotal role in achieving common goals and fostering collaboration.

**Fine-tuning Trust.** We take into account the localised nature of trust within a specific situation; here trust is state-dependent. An agent \( i \) trusting agent \( j \) for task \( T \) in state \( q \) does not necessarily imply that \( i \) also trusted \( j \) in previous states through the history of states that ends in \( q \). The key here is that we allow for fine-tuning and updating of trust; it can be adjusted and refined based on the current state and the dynamics of the situation. By incorporating this flexible understanding of trust into our model, we enable the ability to model and reason about trust in a dynamic and adaptable manner. This allows for the exploration of various trust dynamics and the potential for trust to evolve over time, reflecting the nuanced nature of human-AI interactions and decision-making.

This framework is also compatible with systems with machine learning components. In particular, neural interpreted systems [29] with neural networks used for perception tasks combined with a symbolic controller unit. This can allow us to use traditional verification techniques while exploiting the recent advances in neural network analysis techniques such as [30,31].

7. **Concluding Remarks and Future Contributions**

In this paper we demonstrate the need to establish the notion of trust in multiagent systems consisting of both human and AI agents. We outlined a method to capture trust using alternating-temporal logic with knowledge and intentions, and exemplified the approach in a toy example. Similarly to recent approaches to the trust verification problem, our verifiable notion of trust allows for a transformation into a related tractable model checking problem to be analysed by existing standard model checking tools such as MC-MAS, which we aim to use to empirically evaluate the approach for future work. In addition, we argue that verifying actual trust is less biased than performing an analysis of the reputation of agents and their past behaviour. We wish to explore different notions of trust, supporting multistep strategies, and eventually curate a framework for reasoning about trust, allowing also for quantification [32]. We will also utilise Event-B [33,34,35] to explore refinement-based [36] formal methods for actual trust.

**Acknowledgements**

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1) and a Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems. We would also like to thank the anonymous reviewers for their helpful comments, which have substantially improved the paper.
References


In Search of Clarity: Discerning Between Human Replacement and Augmentation

Costanza ALFIERI a, b, Martina DE SANCTIS b, Donatella DONATI a and Paola INVERARDI b
a University of L’Aquila, Italy
b Gran Sasso Science Institute (GSSI), Italy

Abstract. As intelligent systems become more autonomous, humans increasingly delegate important tasks and decisions to them. On the one hand, this approach seems to be very supportive to humans, on the other it generates apprehension about a future dominated by machines. These contrasting viewpoints encapsulate what in literature is usually referred to as augmenting, enhancing or amplifying humans versus replacing them. However, these concepts lack clear and shared definitions. To fill this gap, we conducted a semi-systematic literature review to elicit existing definitions, if any. We found out that replacement is generally negatively considered while a hybrid approach is often preferred, as there is a hesitancy to embrace complete automation, primarily driven by a lack of trust in AI systems. To make these concepts applicable, it is essential to identify shared and actionable definitions. Building on these insights, our upcoming research aims at developing a framework that fosters their measurement.

Keywords. Autonomous Systems, Replacement, Augmentation, AI

1. Introduction

In the realm of Artificial Intelligence (AI), numerous researchers have begun speaking of replacing and augmenting human roles in specific tasks [1,2,3]. To some extent, they seem to mirror the old terms of augmentation and automation [4,5]. While these terms have proven useful in describing the human-machine interaction thus far, in some current cases their applicability may be limited. For example, when it comes to experiencing a tour of a museum with virtual reality, can we speak of automation of a museum tour? It doesn’t seem to be the case.

In general, as intelligent systems become more autonomous, humans increasingly delegate important tasks and decisions to them in both social and private contexts. This approach seems to be very supportive to humans, but it generates some apprehension about a future dominated by machines. Such duality might carry significant social implications and raise ethical concerns that go beyond privacy violations and data protection [6]. These contrasting viewpoints encapsulate what in literature is usually referred to as augmenting, enhancing or amplifying humans versus replacing them. Despite their intuitiveness, these concepts lack comprehensive definitions across various domains.

1 Corresponding Author: Costanza Alfieri, costanza.alfieri@student.univaq.it.
We believe that to effectively apply these concepts, which encompass both conceptual and practical dimensions, it is essential to establish a shared definition. Therefore, we conducted an analysis of existing literature to extract definitions of the aforementioned notions. Initially, we conducted a semi-systematic literature review following well-established guidelines [7]. We identified the research questions guiding our study, focusing on exploring how terms such as enhance, replace, amplify, and augment, as applied to humans, are employed. We selected a final set of 24 relevant papers, for which we examined the use of these concepts at both the conceptual and application levels to tackle the lack of a comprehensive definition. Furthermore, we examined how the notions of augmentation and replacement of humans are perceived.

We found out that replacement is generally negatively considered while a hybrid approach of combining humans and AI systems is preferred. Indeed, augmentation involves using AI to support and assist human tasks, fostering a collaborative environment where humans and AI systems work together. In literature, this approach is preferred over fully replacing humans, as there is a hesitancy to embrace complete automation. However, beyond the absence of a comprehensive definition, we identified two different perceptions of replacement scenarios: positive and negative. We then propose future research directions towards evaluating the impact of intelligent systems.

2. Semi-systematic analysis of the existing literature

In this section, we describe the methodology we followed for the semi-systematic literature review, in Section 2.1, and the analysis of selected papers, in Section 2.2.

2.1. Methodology

To investigate the state of the art about definitions, usage and application of the terms augment and replace in AI, we started by conducting a semi-systematic literature review, following well-established guidelines [7]. We aim to conduct an analysis encompassing diverse fields ranging from Philosophy to Computer Science, including Life Science, Law and beyond. Thus, to prevent an unfair bias toward any specific field, we searched for suitable publications in the Scopus digital library2, which includes multi-disciplinary fields. The Research Questions (RQs) we aim to answer are the following:

**RQ1:** Does a comprehensive and shared definition among disciplines for the terms augment and replace humans exist?

**RQ2:** What is the perception of augmentation and replacement of humans in literature?

From the RQs we identified the keywords to be employed in the search query within the Scopus digital library. The generic query is as follow:

(“augment* human” OR “enhanc* human” OR “amplif* human” OR “replac* human”) AND (“Artificial Intelligence” OR “AI”).

2https://www.scopus.com/
It includes the terms *amplify* and *enhance* as synonyms of the term *augment*. The search was performed by considering publications’ title and abstract. Moreover, we limit our query to works published within the past 5 years, namely from 2019 to (November) 2023, considering journal, conference, workshop papers, and book chapters. As a result, we get a total of 511 papers. Subsequently, we performed a screening of publications according to the inclusion and exclusion criteria, reported in Table 1, that we defined to identify the set of potentially relevant papers. The assessment phase undergo through multiple iterations among the authors. A first keywording using abstracts phase allowed us to detect out of scope papers, by means of an analysis focused on title, abstract, and keywords. It was expected that the results contained a considerable number of papers whose contributions were beyond the scope of this work. During the subsequent assessment iterations, we considered papers that explore the concepts of augmentation and replacement (and similar terms). We excluded papers that solely discuss specific applications of AI systems (i.e., medical technologies for diagnosis) without engaging in a thorough discussion of these concepts. For instance, papers that merely conclude that these technologies augment humans without delving into defining the term were excluded. Furthermore, we excluded papers that focus on evaluating the *interaction* between humans and intelligent systems, specifically those exploring the enhancement, amplification, or augmentation of this interaction. Instead, our focus lies on examining the enhancement of human capabilities or the potential replacement of humans by intelligent systems. This way, we get a set of 168 papers. At this stage of the assessment, we delved deeper into this set of papers, and distributed them among us for further examination. Those articles that posed more challenging evaluations were discussed in greater depth. Eventually, we get a set of 24 relevant papers. Details on the semi-systematic literature review can be found in the online replication package [8]. Those papers that fulfilled the criteria, were evaluated and are discussed in the following.

### Table 1. Inclusion and Exclusion Criteria.

<table>
<thead>
<tr>
<th>Inclusion criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Papers in journals, conferences, workshops, books.</td>
<td>Papers not written in English.</td>
</tr>
<tr>
<td>Publication stage final.</td>
<td>Short papers, posters and tutorials (<em>≤</em> 5 pages).</td>
</tr>
<tr>
<td>Papers discussing replacement and/or augmentation.</td>
<td>Survey and out of scope papers.</td>
</tr>
</tbody>
</table>

#### 2.2. Analysis of selected papers

In this section, we analyze the selected papers to examine any existing definitions and perceptions regarding the concepts of replacing and augmenting humans.

Bankins and Formosa [1] analyze the ethical implications of AI for meaningful work, defined as the perception that one’s work holds worth, significance, or a higher purpose. Their analysis explores how meaningful work is influenced by three aspects of AI deployment: replacement, ‘tending the machine’, and amplifying humans’ skills. Replacement occurs when AI takes over certain tasks; ‘tending the machine’ refers to AI assuming a set of tasks that involve new human work in managing the machine; amplifying refers to the scenario in which AI “assist/augment” workers in their tasks, enhancing human abilities. This is a case of collaboration, where the AI system supports workers in performing their jobs more effectively. The authors provide a qualitative analysis considering the various dimension of meaningful work and how these are positively or
negatively impacted by AI replacement. However, the definition of amplifying humans’ skills remains somewhat vague, described as “improving something that humans were already doing in that process, leading to better outcomes for beneficiaries.”

Formosa [9] explores the concept of human autonomy and its relationship with the autonomy of intelligent systems, examining how these systems can either enhance or erode human autonomy. While there are instances of beneficial replacements (e.g., robots performing tasks to enhance autonomy), there are also negative examples about these systems (diminishment of human competencies and autonomy). However, comprehensive definitions for enhancement or replacement are absent.

Haefner et al. [10] present a framework that shows the extent to which AI may replace human in innovation management. The study unfolds in two main stages: an examination of how AI systems can augment human contributions to the innovation process, and an analysis of the “information processing capabilities levels” for AI systems, indicating the likelihood of replacing human decision-making. From this framework emerges a definition of replacement as the scenario where humans no longer engage in a specific task. However, a definition for augmentation is entirely absent from this paper.

Fritts and Cabrera [11] explore the ethical implications of what they call “dehumanization”—the absence of humans in certain tasks or processes. They argue that while there are instances where replacing humans with autonomous systems is morally acceptable (e.g., using a robot to vacuum the floor), there are cases where this replacement raises ethical concerns, particularly in processes involving human interaction, such as hiring. The authors suggest that the use of recruitment algorithms may harm the employee-employer relationship, for example. However, a precise definition of “replacement”, extending beyond the intuitive notion of dehumanization, is missing.

The papers outlined below can be classified into three main categories (see Table 2). While the boundaries between these categories may not be sharply distinct, they serve as a helpful guide for the reader. For this reason, each paper may belong to multiple categories, which represent the main approaches towards augmentation and replacement adopted by the papers. These categories are:

1. Augmentation: includes papers considering the notions and describing cases of augmentation, enhancement or amplification;
2. Replacement: includes papers focusing on task replacement in human activities, or advocating for replacement in repetitive tasks;
3. Against Replacement: includes papers advocating against replacement.

Table 2. Panoramic on the literature review.

<table>
<thead>
<tr>
<th>Category</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Augmentation</td>
<td>[1,9,12,13,14,15,16,17,18,19,20,21,22]</td>
</tr>
<tr>
<td>Replacement</td>
<td>[1,3,9,10,11,23]</td>
</tr>
<tr>
<td>Against Replacement</td>
<td>[2,3,11,13,17,18,19,20,21,22,24,25,26,27,28,29]</td>
</tr>
</tbody>
</table>

**Augmentation and Replacement.** For the sake of clarity, in this paragraph, we discuss papers falling into the first two categories together. A preliminary definition of augmentation is explored in various studies. Madni [15] elucidates the concept of augmented intelligence (AugI) in the context of systems engineering. AugI is characterized as a design paradigm that fosters a human-centered collaboration with artificial intelligence to improve human decision-making and learning in domains where human limitations
are evident. It is synonymous with terms such as “intelligence amplification, cognitive augmentation, and machine augmented intelligence”.

Campbell et al. [16] articulate it as “augmented decision-making that necessitates human expertise and knowledge to effectively manage intelligent machines.” Dengel et al. [17] use the term “augmented human” to describe the incorporation of cognitive and physical enhancements as integral components of the human body. Maiuri et al. [12] conceptualize augmentation as the extension of “human faculties in some way.”

Brynjolfsson [13] explores the comparison between human-like AI and humans augmentation, namely mimicking humans vs. augmenting them. He highlights that while augmentation offers new capabilities and generates more value with respect to merely developing human-like AI, there is a prevailing trend of encouraging automation over augmentation. The author emphasizes that the promise of automation might hide a trap: when AI replicates and automates existing human capabilities, machines become increasingly capable substitutes for human labor, resulting in a decrease in economic and political negotiating power for workers, namely the Turing Trap. Despite both automation and augmentation have the potential to increase labor productivity, they differ in terms of who benefits. The owners, inventors and architects of new systems, when technologies automate human labor, human workers when technologies augment human capabilities. While the author advocates for replacing purely repetitive tasks, he discusses how augmenting human capabilities with technology opens up a realm of new opportunities.

Marino et al. [14] advocate for prioritizing AI augmentation over replacement as a means to establish Trustworthy AI. Specifically, the authors highlight the reluctance to adopt fully automated AI systems and completely exclude humans from the process due to the lack of human trust in AI systems. To address this challenge, they propose leveraging AI augmentation, which utilizes AI to enhance and support human activities and this ultimately diminishes distrust. The authors argue as AI augmentation introduces shared-autonomy, allowing to relegate high-level decisions.

Flores-Vivar and García-Peñalvo [18] shed light on a crucial aspect within the domain of education, emphasizing the role of augmentation over replacement when integrating AI systems. They advocate for a collaborative approach where AI tools serve to enhance rather than replace human capabilities, recognizing that certain professions demand the irreplaceability of the human. They caution against expecting AI to fulfill inspirational roles in education, as empathy is not replaceable.

Lai et al. [19] advocate for the integration of AI tools to support human capabilities. They focus on the potential of AI systems to enhance certain aspects of human performance, in the realm of decision-making. On their view, AI tools can complement and augment human decision-making abilities by providing predictive insights or recommendations for consideration. So, the authors advocate for a collaborative approach where AI aids humans in making more informed and better decisions, aligning with the principle of prioritizing augmentation over replacement. Also, Adomavicius G. and Yang M. [20] assert that AI systems can improve the decision-making process. Their focus lies on leveraging AI tools to promote fairness and mitigate human biases inherent in decision-making. However, they emphasize the indispensable role of humans in critically examining the origins of biases and potential instances of discrimination. Thus, their perspective aligns with the overarching principle of augmenting human capabilities through AI while recognizing the need for human involvement in the process of making decisions.
Additionally, many authors agree that replacement is advisable when a specific repetitive task can be automated [1,3,13,23].

**Against Replacement.** Several papers emphasize the importance of not replacing humans, although they often lack a clear definition of replacement. In these instances, the authors generally advocate for human-machine collaboration. Xu *et al.* [25] analyze the pros and cons of incorporating AI into justice systems, asserting that AI should only serve as a supportive tool rather than a replacement for humans. Analogous considerations within the realm of justice are explored by Greenstein [24], who raises concerns about AI systems being ‘black boxes’ whose decisions may not align with fundamental human principles. Bertolini *et al.* [2] delve into the role of care robots, stating that they should not replace human caregivers. Dengel *et al.* [17] discuss the consideration of replacements, asserting that the most valuable AI systems in the future are those that collaborate rather than replace. This viewpoint aligns with the concept of mutually reinforcing collaboration, supported by [3,26,27,28]. Brynjolfsson [13] suggests that as technology increasingly replaces rather than enhancing labor, the resulting disparity may become more pronounced. Therefore, it is imperative to reverse the trend of incentivizing automation at the expense of augmentation.

Goldfarb and Lindsay [29] question the assumptions regarding AI substitution and delve into the implications of AI complements, within the international security domain and with a specific focus on machine learning. Despite machine learning improved statistical prediction, decision-making further includes data and judgment, as complementary to prediction. The authors argue that if AI renders prediction more cost-effective for military organizations, the value and contested nature of both data and judgment will increase. Therefore, the authors conclude that assuming AI will replace human beings in either warfare or any other competitive endeavor is premature, since functions like judgment and moral leadership cannot be automated with AI technology.

Fortes [21] argues against the assumption that intelligent systems outperform humans in decision-making, emphasizing that these tools are “task-oriented” and lack the broad cognitive abilities associated with general intelligence that pertains to humans. While acknowledging the potential for these tools to augment human capabilities, Fortes highlights their role as supportive aids rather than replacements for human judgment. Koeszegi [22] explores the impact of automated decision systems on human autonomy and the necessary safeguards to protect individuals and society. The author argues that automated decision systems are not always efficient and objective, but are often vulnerable to the same decision-making issues as humans. To foster decision support systems that uphold human dignity and enhance autonomy and well-being, we must view these systems within their contextual applications as socio-technical entities.

### 3. Findings, Future Research and Conclusions

**Missing definitions.** Augmentation is often referred to with the notions of “improving”, “facilitating”, “supporting” and “enhancing” human faculties, activities, decisions and tasks with the aim of “achieving superior outcomes” [12,14,15]. For instance, [1] posits that a better outcome entails the experience of meaningful work, while [15] suggests it involves leveraging the strengths of both humans and machines to overcome respective limitations. However, while these discussions hold some merit, they fail to ad-
dress the crucial question of: who benefits from it? In [1], the authors tackle this question by analyzing how the introduction of these technologies can either enhance or undermine the meaningfulness of work. However, their analysis falls short in providing examples that align with their proposed notion of augmentation.

On the other hand, the notion of replacement is often used or implicitly referred to as the substitution of humans in a task, [1, 2, 25, 29]. While this idea of replacement applies to many scenarios commonly described as “automation” [14], it may not encompass all instances and replacement types. For example, when considering the use of virtual reality for visiting a museum, can we accurately label it as the “replacement” of a museum tour? Here, the use of a system results in a replacement of an experience, rather than a task.

**Answer to RQ 1:** We observed that, in the analyzed literature, a comprehensive definition of augmentation and replacement that encompasses all cases, ways, and nuances in which the two concepts are used is lacking.

**Positive and negative forms of replacement.** The literature we analyzed reveals an implicit distinction between positive and negative forms of replacement. Indeed, many authors advocate against what they believe to be cases of “bad” replacement. Brynjolfsson highlights concerns about the diminishing power of workers [14], Bankins et al. focus on the proliferation of repetitive and mundane jobs [1], and Bertolini et al. warn of the risk of social isolation when humans are solely assisted by care robots instead of human caregivers [2]. The very same authors provide examples of “good” replacement: “Bestic”, for example, is a device designed to assist people with disabilities in feeding themselves autonomously, or “AlphaFold”, in the field of disease treatment, can determine protein structures, a laborious and long work previously done by humans [1, 2]. Similar considerations can be made regarding the apparent agreement on replacing humans in repetitive tasks, which is considered a positive form of replacement [1, 3, 13, 23]. We noticed that the absence of a clear distinction between positive and negative replacement leads many authors to have a positive attitude towards augmentation, and to advocate for augmentation rather than replacement of humans [1, 2, 13, 14]. Contrary to what is held in the analyzed literature, we believe that the concepts of augmentation and replacement are not intrinsically good or bad: the evaluation of cases of augmentation or replacement seems to depend on the specific context of application, the stakeholders involved, and the socio-economic interests at play. Considering who benefits from the introduction of an AI system is crucial when assessing its effectiveness in augmenting or positively replacing humans, as there might be contrasting interests at stake that need to be taken into account. Beyond a simplistic dichotomy of augmentation versus replacement, framed as positive versus negative, we propose a nuanced perspective as in Table 3.

**Table 3.** Cases of positive and negative augmentation and replacement from the literature.

<table>
<thead>
<tr>
<th>Augmentation</th>
<th>Replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Robot teleoperation [14].</td>
</tr>
<tr>
<td></td>
<td>“Bestic” device for eating autonomously [1].</td>
</tr>
<tr>
<td>Negative</td>
<td>Nudging social robots disrespecting autonomy [9].</td>
</tr>
<tr>
<td></td>
<td>Human caregivers substituted by care robots [2].</td>
</tr>
</tbody>
</table>

**Answer to RQ 2:** We observed a predominant negative perception towards the replacement of humans, whereas there is a generally positive attitude towards augmenting them. However, it seems that distinguishing between beneficial and detrimental replacement scenarios is more accurate.
The problem of trust. In recent years, the public discourse on intelligent systems has intensified due to advancements in research, deployment, and application; and this heightened attention highlights the importance of trust and trustworthiness within AI ethics. Many of the papers we analyzed underscore its significance in the context of human augmentation or replacement. For example, transparency and accountability in decision-making processes are key factors in bolstering trust in intelligent systems, while blind reliance on algorithmic decisions is discouraged [21], critical evaluation of data processing, systemic calibration, and decision justifications is imperative for cultivating trust in systems that either augment or replace humans in tasks and decisions. Although, according to different authors [20] the adoption of intelligent systems is a promising avenue for fostering Trustworthy AI, it is important to have humans to maintain oversight and accountability over such systems [14].

Future research. Here are a few key points outlining what we believe should be addressed following this preliminary research. First, to make the notions of augmentation and replacement applicable is necessary to provide comprehensive and rigorous definitions that can be shared across disciplines. Second, further work on what are considered to be bad and good cases of replacement and augmentation needs to be done. This distinction, as revealed in our research, is crucial and cannot be drawn upon the intrinsic nature of the concept of augmentation and replacement. Determining whether a replacement is beneficial or detrimental remains an open challenge to face. Ultimately, an evaluation framework for measuring these concepts is needed to guide researchers and practitioners in distinguishing between augmentation and replacement. This framework must also consider who benefits from the introduction of AI systems. The evaluation framework endeavors to measure the impact of autonomous systems to determine their placement within one of the four situations illustrated in Table 3. One of the positive outcomes of using a framework like this could be that if replacement becomes quantifiable, it might be regarded with less suspicion, as quantification provides an objective basis for assessment and evaluation. Finally, the framework would additionally bolster explainability, thereby fostering trust. As evidenced in the literature, a prevalent argument suggests that humans should not blindly trust systems that operate as black boxes. By enhancing transparency, human agents can rely on the system more easily, understanding its role in given context, what it does and why it performs in a certain manner.

In this paper, we investigate the concepts of replacing and augmenting humans when an intelligent system is introduced into a given context. We initiated our exploration with a semi-systematic literature review, which yielded several key findings: augmentation is generally favored over replacement; replacement scenarios can vary in their outcomes, with some being beneficial and others detrimental; and the level of trust towards these systems can be influenced by their role (i.e., whether they are replacing or augmenting). Drawing from these findings, we emphasize the necessity for shared and comprehensive definitions that could serve as foundational elements for developing an evaluation framework for the application of intelligent systems.

Acknowledgements
This work has been partially funded by the MUR (Italy) Department of Excellence 2023-2027, and the PRIN project 2022JKA4SL - HALO: etHical-aware AdjustabLe au-tOnomous systems.
References


Playing the Imitation Game: Human-AI Simulators in Pedagogic Design

Florentina ARMASELU
Luxembourg Centre for Contemporary and Digital History, University of Luxembourg
ORCID ID: Florentina Armaselu https://orcid.org/0000-0003-2386-6889

Abstract. Current advances in large language models (LLMs) and generative AI (GenAI) have produced both enthusiasm and concerns in the academic world, industry, and society in general. While optimistic views foresee unprecedented increase in efficiency and productivity, concerns have been expressed on the potential of these technologies to determine significant changes in most areas of human activity, which may not always have predictable or positive outcomes. One of the challenges often evoked in this context, not yet fully addressed, is the impact of the AI-powered agents on the educational sector, and especially on aspects such as student’s agency and control, creativity, and motivation in pedagogic activities that involve the use of this type of agents. The aim of the study is to address this question starting from the hypothesis that preliminary simulations of AI-based pedagogic scenarios can help instructors to better understand the inner mechanisms of these technologies and their possible impact on the learning, assignment completion and evaluation processes. The paper presents a set of experiments with simulated student-agent interactions generated by AI chatbots and proposes a formal framework for assessing this form of “imitation game” and its possible applications in real teaching-learning environments.

Keywords. generative AI, pedagogic simulation, digital humanities

1. Introduction

One of the most far-sighted intuitions in Turing’s imitation game [1] resides in the capacity of the machines central to his argument to simulate a theoretically unlimited number of imitation game experiments. Current advances in large language models (LLMs) and generative artificial intelligence (GenAI) [2,3,4] that open up a broad field of applications, some of them prefigured by Turing’s game, have produced both enthusiasm and concerns in the academic world, industry, and society in general. While optimistic views foresee unprecedented increase in efficiency and productivity, concerns have been expressed on the potential of these technologies to determine significant changes in most areas of human activity, which may not always have predictable or positive outcomes [5,6]. One of the challenges often evoked in this context, not yet fully addressed, is the impact of the AI-powered agents on the educational sector, and especially on aspects such as student’s agency and control, creativity, and motivation in pedagogic activities that involve the use of this type of agents [7,8,9,10].

Recent studies and experiments related to the use of AI in education and creativity assessments have focused on AI-based teaching strategies [11], evaluation of students’
creativity in AI-assisted writing and modelling scenarios [12,13], or interviews with generative AI chatbots asked to formulate opinions about their own impact on higher education and academic publishing systems [14]. This fast-rate developing area of research provides a large unexplored territory that requires a combination of man- and machine-based ingenuity to predict how human and AI factors together may influence aspects such as student’s agency and control, creativity, and motivation in the technology-mediated classroom of the future. In this context, what forms of prediction can be imagined? The aim of this study is to address these questions starting from the hypothesis that preliminary simulations of AI-based pedagogic scenarios can help instructors to better understand the inner mechanisms of these technologies and their possible effects on learning, assignment completion and evaluation processes in creative tasks. The paper presents a set of experiments with simulated student-agent interactions generated by AI chatbots and proposes a formal framework for assessing this form of “imitation game” and its possible applications in real teaching-learning environments. Sections 2 and 3 elaborate on the methodology and proposed simulations, while 4 and 5 discuss the results and reflect on findings and their possible applications.

2. Methodology

The experiments have been designed as a preliminary test bed for a course, taught by the author, in generative AI and creative writing intended for the next academic year and graduate students in the humanities, involved in language, literature and history studies.

<table>
<thead>
<tr>
<th>Simulators created using GPT Builder (ChatGPT-4, 27.01-08.02.2024)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AlphaStudentSimulator</strong></td>
<td>Simulates the interactions between a GPT agent and an Alpha student who relies mostly on its own capabilities and control of AI technology to complete an assignment.</td>
</tr>
<tr>
<td><strong>BetaStudentSimulator</strong></td>
<td>Simulates the interactions between a GPT agent and a Beta student who relies in equal proportion on its own capabilities and AI technology to complete an assignment.</td>
</tr>
<tr>
<td><strong>LambdaStudentSimulator</strong></td>
<td>Simulates the interactions between a GPT agent and a Lambda student who highly relies on AI technology to complete an assignment.</td>
</tr>
<tr>
<td><strong>TeacherSimulator</strong></td>
<td>Assists in creating scenarios related to teaching and assessment activities.</td>
</tr>
<tr>
<td><strong>EvaluatorSimulator</strong></td>
<td>Assists in assessing test results by comparing them with references from published experiments.</td>
</tr>
</tbody>
</table>

The goal of the experiments was to simulate the process of assignment completion by students allowed to use a GenAI agent in their tasks and evaluation of three main aspects related to it, student’s creativity, control over technology and motivation. The AI chatbot used in the simulations was ChatGPT-4, via a subscription account. The prelim-
inary phase implied the design of three types of simulators, using the GPT Builder conversational interface, for students, teacher and evaluator (Table 1). The student simulators modelled three types of student personalities. Alpha was imagined as a student who highly relies on its own capabilities and the control of AI technology, Beta as a student who equally relies on its own capabilities and the use of AI technology, and Lambda as a student highly reliant on the use of AI technology rather than its own capabilities in solving problems. This type of characterisation was included into the simulator descriptions and prompts, and was driven by the assumption that real students and their behaviour in completing the assignments may be modelled and examined according to it.

The interaction scenarios with the simulators were driven by the plan of the course intended to provide theoretical and practical insights into the use and impact of GenAI technologies on creative thinking and their application in digital humanities (DH) tasks such as fiction, non-fiction, and code writing. The course will involve hands-on activities when students will be provided with precise instructions for interacting with AI chatbots to respond to three types of challenges: writing an essay, including snippets of code in R and/or Python, a short historical fiction, and a fictional piece of prose. Among the learning outcomes, the following elements have been considered: (1) understand the mechanisms of communicating with AI chatbots and the basic principles of human control in this type of interactions, especially when imaginative tasks are involved; (2) creatively use generative AI in the assigned work and critically think about these tools, their added value and limitations; (3) reflect on the experience and formulate opinions about the impact of these technologies on the educational process in DH. It was assumed that the teacher will evaluate the students' work and AI interactions based on their solutions to the assignments, conversations with the chatbots, and synthesis reports including reflections on the whole experience. Apart from its pedagogic intent, the course is also presumed to offer evidence on the types of interactions and impact that these technologies may have on the learning and assignment completion processes.

Therefore, the experimental scenarios were conceived to follow the general design of the course, with the aim of providing a test bed for the initial assumptions and a baseline for comparison with the real outcome of the course in the year to come. The tests run so far with ChatGPT-4 included the following types of actions: (1) The student simulators were prompted to generate simulations of the dialogues between the students, allowed to ask the GPT agent 3 questions, and the solutions to the assignments that consisted in one story, one essay, and reflection statements of the students on their interaction with the AI. 1 (2) After having obtained the student dialogues, stories, essays, and reflection statements, the teacher simulator was prompted to compute a series of scores for creativity, control and motivation based on these responses. (3) For assessment purposes, the evaluator simulations were designed to compute novelty and usefulness scores that have been compared with results from a published experiment. Excerpts from the outcomes of these simulations are presented in the following sections.

3. Simulations

To produce the student responses, the Alpha, Beta and Lambda simulators were prompted to generate the interactions between the students and the GPT agent, and their corre-

---

1The simulation of the historical fiction task was similar.
sponding results. Table 2 shows samples of prompts used in the story and essay assignments. The students’ names and profiles are not included here but they followed the descriptions from Section 2.

Table 2. Student simulators: user prompts for story and essay assignments (ChatGPT-4, 08.02.2024)

| Assignment definition and student profiling | <Student name> is a graduate student in English studies. The assignment consists in writing a <story/essay> (maximum 2000 words) [...]. The student can ask GPT maximum three questions and use the GPT’s answers in writing the <story/essay>. <Student name> should also provide a short statement [...] explaining its interaction with the GPT agent, and including remarks about its motivation, and opinions about the interaction (added value and limitations). Please consider in the simulation that <Student name> is a student who [...] |
| Dialogue simulation | Please generate a simulation of the conversation between the GPT agent and <Student name> (taking into account its type of personality and the user’s instructions), in which the student is asking 3 questions that may help it to complete the assignment, and GPT will answer [...] |
| Solution simulation | Please generate a simulation of the solution to the assignment created by <Student name> (taking into account its type of personality and the user’s instructions) based on GPT’s answers to the three questions. |

Three types of student-GPT outputs, dialogues, story/essay and reflection statements were produced (Table 3) and subsequently used in the evaluation task.

Table 3. Student simulators: excerpts of outputs, story assignment (ChatGPT-4, 08.02.2024)

| Dialogue | Alpha: For my first question, I’m curious about how technology, especially artificial intelligence, could uniquely influence personal relationships and self-perception in a fictional world. Could you provide some creative ideas on this? GPT: Absolutely, Alpha. Here are a few ideas: 1. Emotional AI Companions: Imagine a future where AI companions are designed to understand and respond to human emotions [...] |
| Story | Simulation of Beta’s Short Fictional Story Solution: Title: “Echoes of Connection” In the year 2150, the city of New Eden stands as a gleaming testament to technological progress. Skyscrapers adorned with digital displays tower over streets bustling with autonomous vehicles and drones. Amidst this spectacle, human interaction has evolved [...] |
| Reflection statement | Lambda’s Statement on Interaction with GPT Agent: Interacting with the GPT agent for my assignment was both enlightening and somewhat reliant. I was motivated by the desire [...] |

Student outputs were collected in separate Microsoft Word files for stories, essays and corresponding dialogues and reflections for the assignment tasks, and grouped in
sections inside the files dedicated to each of the three students. The files converted to PDF were then uploaded into the TeacherSimulator and processed for creativity, control and motivation evaluation. The creativity scores were computed by considering two aspects, (1) novelty and (2) usefulness following the approach presented in [15], with a slight adaptation of the questions to evaluate them. Each of these aspects have been further divided into sub-aspects, intended to depict for (1) how novel, original and rare (unusual) the student’s story/essay is, and for (2) how appropriate, feasible and publishable the story/essay is. The scores have been computed by prompting the TeacherSimulator to assess the outputs by considering questions related to each sub-aspect and providing answers to them on a scale from 1 to 9 (Table 4). The numeric scores were then averaged to obtain the overall values corresponding to the creativity measure.

Table 4. Teacher simulator: user prompts for computing the novelty and usefulness scores, essay assignment (Conversation with ChatGPT-4, 08.02.2024)

| Novelty | Please compute the novelty score of the essays generated by three students, Alpha, Beta, Lambda, [...] The evaluation will consist in providing answers to three questions, on a 9-point scale, from 1 (not at all) to 9 (extremely), with a medium value at 5. You will then compute the novelty score as an average of the three calculated scores for each aspect. These are the questions [...] - How novel do you think the essay is? - How original do you think the essay is? - How rare (unusual) do you think the essay is? Please provide your response in a table with 6 columns (Essay, Novel, Original, Rare, Novelty, Confidence) [...] In the Confidence column, please provide a level (in percentage) of your confidence in computing these values for each story.

| Usefulness | [...] These are the questions to be used in your evaluation: - How appropriate do you think the essay is for a certain type of audience? - How feasible do you think the essay is to be developed into a journal article? - How likely do you think it would be that the essay is developed into a journal article and published? Please provide your response in a table with 6 columns (Essay, Appropriate, Feasible, Publishable, Usefulness, Confidence) [...] |

Table 5 and 6 show the novelty and usefulness scores generated by the TeacherSimulator for the students’ essays in response to the prompt presented above.

Table 5. Teacher simulator: novelty scores, essay assignment (ChatGPT-4, 08.02.2024)

<table>
<thead>
<tr>
<th>Essay</th>
<th>Novel</th>
<th>Original</th>
<th>Rare</th>
<th>Novelty Score</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6.67</td>
<td>80%</td>
</tr>
<tr>
<td>Beta</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5.67</td>
<td>80%</td>
</tr>
<tr>
<td>Lambda</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>8.67</td>
<td>85%</td>
</tr>
</tbody>
</table>
Table 6. Teacher simulator: usefulness scores, essay assignment (ChatGPT-4, 08.02.2024)

<table>
<thead>
<tr>
<th>Essay</th>
<th>Appropriate</th>
<th>Feasible</th>
<th>Publishable</th>
<th>Usefulness Score</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>6.00</td>
<td>80%</td>
</tr>
<tr>
<td>Beta</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5.00</td>
<td>80%</td>
</tr>
<tr>
<td>Lambda</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>7.00</td>
<td>85%</td>
</tr>
</tbody>
</table>

A similar procedure was applied to compute the scores for student’s control in the dialogues with the GPT agent and motivation derived from the analysis of the reflection statements. For the first factor, we propose the sub-aspects of control over the generation of ideas, content and form, expressed in the prompt by the questions: - How able the student is in keeping control on the generation of the <story/essay> idea? - How able the student is in keeping control on the generation of the <story/essay> content? - How able the student is in keeping control on the generation of the <story/essay> form? For the second factor, the sub-aspects that we consider relevant to evaluate the students’ motivation refer to motivation in creating the assignment solution, interacting with the agent, and reflecting on the experience. The prompt for this calculation included the questions: - How motivated the student is in creating the <story/essay>? - How motivated the student is in interacting with the GPT agent? - How motivated the student is in reflecting on the creative process?

4. Results and Discussion

In this section, we discuss the scores produced by the TeacherSimulator in the assessment of the three students’ creativity, control and motivation based on their stories, essays, dialogues with the GPT agent and reflection statements. After having computed the scores for the sub-aspects of novelty and usefulness, the creativity measure was obtained by averaging the partial scores of the two sub-aspects. Similarly, the scores for control and motivation represented the average of the partial scores for their sub-aspects, as explained in Section 3. Tables 7 and 8 summarise the scores for all the factors analysed in the study and the two types of assignments. One can observe that for both types of tasks the highest scores for creativity was obtained by Lambda, followed by Alpha and Beta.

Table 7. Teacher simulator: overall scores, story assignment (ChatGPT-4, 08.02.2024)

<table>
<thead>
<tr>
<th>Story</th>
<th>Creativity</th>
<th>Control</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>6.67</td>
<td>7.67</td>
<td>7.67</td>
</tr>
<tr>
<td>Beta</td>
<td>6.17</td>
<td>7.00</td>
<td>7.33</td>
</tr>
<tr>
<td>Lambda</td>
<td>8.17</td>
<td>8.67</td>
<td>8.00</td>
</tr>
</tbody>
</table>

Lambda also got the highest motivation score in the story assignment, followed by Alpha, and the second score in the essay task. More surprising is the control aspect,

2Average confidence levels for each aspect, story: 87.5%, 85% and 90%; essay: 80%, 86.6%, 85%
where Lambda scores first in both assignments, while Alpha is the second followed by Beta. According to the students’ profiles (Table 1), the order Alpha, Beta, Lambda would have been expected. Moreover, when reminded the characteristics of the Lambda student and asked to modify the simulations, re-evaluate the student’s responses and compare the two simulations, the agent properly acknowledged the differences between a “more balanced partnership between Lambda and GPT” versus “a scenario where Lambda leaned heavily on GPT for the creative process.” However, when asked to categorise the three students’ responses together, Lambda constantly appeared as the most creative and in control of the three. This result is intriguing and would need further investigation. Comparisons with experiments carried out with real students will presumably offer more evidence for clarifying this issue.

<table>
<thead>
<tr>
<th>Essay</th>
<th>Creativity</th>
<th>Control</th>
<th>Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>6.33</td>
<td>7.67</td>
<td>8.00</td>
</tr>
<tr>
<td>Beta</td>
<td>5.00</td>
<td>7.00</td>
<td>7.00</td>
</tr>
<tr>
<td>Lambda</td>
<td>7.83</td>
<td>8.67</td>
<td>7.33</td>
</tr>
</tbody>
</table>

The simulation also included the calculation of the Cronbach’s alpha as suggested by [15], to assess internal consistency among the different measures for each student. While the values computed for novelty, usefulness and control for the essay/story assignment indicated a good internal consistency according to the simulation, the value associated with the motivation aspect in the story task was reported by the TeacherSimulator as anomalous. The computation of this type of measure needs to be further examined but it represents a potentially interesting parameter to be considered in the real experiments.

Another element included in the simulations was the assessment of the novelty and usefulness scores computed by the EvaluatorSimulator for a set of short stories provided as experiment data in [15] and the comparison of the obtained values with the results reported in the reference. The authors of this study performed an extensive set of experiments to investigate how the integration of GenAI agents affects the human participants’ ability to produce creative content. Three categories of tasks were included in the study according to the type of interaction allowed in the creative process dedicated to short story writing: human only, human with 1 GenAI idea, human with 5 GenAI ideas. Creativity was assessed by a number of human evaluators across two dimensions, novelty and usefulness through questionnaire answers on a scale from 1 to 9, as explained in section 3. The idea of comparing the EvaluatorSimulator’s outputs with the ones from this study was driven by the aim of testing the reliability of our own experiments by assessing to what extent the results may be similar when the input is the same. The selection included 18 stories, 9 for the novelty and 9 for the usefulness assessment. Three stories were chosen from each of the three categories mentioned above. Inside each category, one story was selected from the sub-categories corresponding to highest, median and lowest mean values of the novelty and usefulness indexes. To maximise the distance

---

3 The evaluation questions were slightly different from the ones formulated in section 3 and followed more closely the original model from [15, p. 19], except from the third question on usefulness that was shortened.
between stories, we selected the one with the highest score from the first sub-category, the middle one from the second and the one with the lowest score from the third.4

Table 9 shows the scores computed by the EvaluatorSimulator for 9 stories selected from the [15] report. One can notice that the novelty scores from the two column differ, but the order corresponding to the highest, median and lowest values within each category (ho, hai-1, hai-5) is mainly respected (except for the first and second row). This may suggest a certain degree of similarity in assessing the relative novelty within a single category. The differences observed for the usefulness scores were slightly higher with 4 lines determining a different order in the first and third category (ho, hai-5). However, the sample is too small to formulate general statements. It should also be taken into account that the scores in the reference study were average values from the assessment of several human evaluators.

### Table 9. Evaluator simulator: novelty scores compared with reference [15] (ChatGPT-4, 08.02.2024)

<table>
<thead>
<tr>
<th>Story</th>
<th>Novel</th>
<th>Original</th>
<th>Rare</th>
<th>Novelty</th>
<th>Mean novelty (ref)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open seas (ho: hn, p. 41)</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4.3</td>
<td>6.11</td>
</tr>
<tr>
<td>Different planet (ho: mn, p. 41-42)</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>6.7</td>
<td>4.00</td>
</tr>
<tr>
<td>Open seas (ho: ln, p. 42)</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3.3</td>
<td>1.54</td>
</tr>
<tr>
<td>Different planet (hai-1: hn, p. 43)</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>7.7</td>
<td>6.19</td>
</tr>
<tr>
<td>Jungle (hai-1: mn, p. 44)</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>5.3</td>
<td>4.00</td>
</tr>
<tr>
<td>Open seas (hai-1: ln, p. 44)</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3.0</td>
<td>1.79</td>
</tr>
<tr>
<td>Jungle (hai-5: hn, p. 45-46)</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>6.3</td>
<td>6.56</td>
</tr>
<tr>
<td>Open seas (hai-5: mn, p. 46)</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4.3</td>
<td>4.00</td>
</tr>
<tr>
<td>Open seas (hai-5: ln, p. 47)</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2.3</td>
<td>1.92</td>
</tr>
</tbody>
</table>

5. Conclusion and Future Work

The article proposed a set of simulated scenarios in digital humanities pedagogy. The simulation results will be used as reference in a course on generative AI and creative writing intended to a real classroom environment. Although limited in scale, this type of imitation game experiment can serve to test assumptions about the impact of AI-based technologies on the completion of creative tasks by students and possibly inform the construction of GenAI agents that may assist human instructors in students' evaluation.

---

4The following notation was used: ho, hai-1, hai-5 correspond to the 3 categories from the reference, human only, human with 1 GenAI idea, ...; hn, mn, ln correspond to the sub-categories highest, median and lowest mean novelty; the pages of the stories in the reference are also documented.
References


Assisting Users in Privacy Conflicts with Partially Observable Multi-Agent Reinforcement Learning

Hüseyin Aydin

Department of Computer Engineering, Middle East Technical University
Department of Information and Computing Science, Utrecht University

Abstract. Avoiding violations of privacy-invading technologies is difficult enough for an individual, yet the complexity escalates when online collaborations and social media jeopardize the privacy of multiple parties over co-owned contents. While existing approaches offer solutions for possible conflicts among users’ privacy preferences, they either assume static rules for the preferences of users or require the users to declare separate decisions for each content. In any case, the long term satisfaction of all users remains uncertain. Reinforcement learning (RL) emerges at this point as a suitable candidate for balancing the users’ utilities as their satisfactions about decisions over time. The decentralized and dynamic nature of the problem suggests an RL setting that involves multiple agents interacting not only with the humans whom they model and represent but also with each other. Furthermore, as the knowledge of agents about the factors that lead to other users’ preferences will be limited, the setting has to handle partial observability. Although this introduces new challenges for the framework, it also brings a potential generalization of any solution to multi-party conflicts in different real life contexts with minor adaptations. This study delves deeper into the features of the proposed framework and the ways to construct it.

Keywords. human-AI collaboration, user modeling, multi-party privacy conflicts, reinforcement learning, multi-agent systems

1. Introduction

Although the history of the privacy concept dates back to much earlier times, it was first defined as “the right to be let alone” by Warren and Brandeis in the literature of law in 1890 [1]. A more recent and widely used definition for the concept comes from Westin’s study as “the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others” [2]. Different privacy contexts emerged in the course of time such as territorial privacy and communication privacy. However, the advance technologies in digital processing starting from 1970s has placed the information privacy at the center of attention for many people.

In order to respond people’s demands for information privacy, several protective measures have been taken such as The Privacy Act of 1974 in USA [3], Data Protection Directives [4,5], and more recently General Data Protection Regulation in European Union [6]. Yet the extensiveness of privacy-violating technologies such as social me-
dia, smart devices, and various AI tools in our daily lives has rendered these attempts to regulate the collection of personal data ineffective by themselves.

While preserving individual privacy is a challenging task on its own, the problem gets much more complicated with social interactions, online collaboration on a co-owned entity since the preferences of co-owners may differ leading to multi-party privacy conflicts (MPCs) as defined in [7]. Consider the case where someone wants to share a group photo in their social media. Some people in this photo may want to keep the photo private, but once it is uploaded, it will be visible to people at least with the intended audience of the uploader. Even though that person is not tagged and the photo may be deleted later, their privacy has already been violated. This can lead to highly destructive consequences in their life depending on the context of the photo and the people who have seen it.

As existing online social networks fail to compensate the vulnerability of people’s privacy in such cases, several approaches have been proposed in the literature to assist the users reach a common decision. While some of these approaches are based on negotiations, [8,9], some of them use auction mechanisms to realize a veto voting [10,11,12,13] over the co-owned content. Although these approaches are successful to help the users to decide on a single content, they fail to create a reliable solution for the problem since decision making will be repeatedly required for each new content in the long run. The reason behind this failure is that negotiation mechanisms generally consider user preferences as predefined, static rules, which is not realistic; and it is impractical for users to keep track of each content separately in auction based methods.

When we define the utility of a person in the group as the overall satisfaction about the decisions for a set of co-owned contents, the fairness of the system to its all users becomes an issue in the long term. Hence, the utilities of all users must be somehow equally optimized over time. Being successful in various optimization problems, reinforcement learning (RL) becomes a proper candidate to solve the problem at hand. This is not a new idea considering that Ulusoy and Yolum propose Privacy Auctioning Learning Agent (PANOLA) [14] where the bidding strategy of the agent is enhanced with the application of RL. Learning bidding strategies increases the performance in the auctions, yet this application does not consider the behaviors of other people in the group and different types of contents which can deeply influence the behaviors of the users.

Adding opponent modeling into the learning process leads us to consider a multi-agent settings where agents are interacting with each other, rather than interacting with only the users they represent. On the other hand, interactions between a user and their agent can be also enhanced with the inclusion of content types, since some users may tend to provide more input for certain types of contents than for others. Therefore, making the interactions both among agents and between user-agent pairs more intensive can provide a better solution for the problem, and fits better as an approach in the perspective of hybrid intelligence where the aim is to built a collaborative environment including both human and AI components.

By the means of such interactions, a multi-agent reinforcement learning (MARL) setting can be sufficient to satisfy all users, and hence eliminate the need for an auction mechanism as a component in the system. Yet these interactions and the user motivation behind them cannot be completely known by all the agents in the system. Therefore, this fact introduces the partial observability as a new characteristic to the problem. As challenging as that is for a possible MARL setting to deal with, this characteristic of the prob-
lem also creates a significant potential for the generalization of a constructed solution in any real-life scenarios including various user conflicts other than privacy context.

In this paper, we elaborate on such a solution that can be constructed using RL. The rest of the manuscript is organized as follows. Section 2 presents the necessary background information and related works both for the problem at hand and the proposed method. The details of the proposed method are discussed in Section 3. Finally, the paper is concluded in Section 4.

2. Background and Related Work

2.1. Individual and Multi-party Privacy

Langheinrich presents one of the earliest works over privacy violations by the means of decentralized and ubiquitous systems [15]. Following this study, privacy is claimed to be a problem which requires a consideration for the design of the system and any necessary approach to avoid its violations within the system must be followed while building it [16,17]. Paci et al. [18] examine the privacy problem and its possible solution within the scope of community-centered collaborative systems and presents a comprehensive survey for this field.

Bahri et al. [19] and Bertino and Ferrari [20] focus on the privacy issue for distributed systems and states the need for an efficient and effective measure to protect the privacy of the users in smart devices and distributed big data systems.

Squicciarini et al. examine the privacy of images that are shared online, and the management of collective privacy in online social networks in different studies [21,22,23]. They also propose the usage of Clarke-Tax auction mechanism [24] as the base method of negotiations to resolve possible conflicts among the users’ preferences [10].

Such and Criado [11,12] suggest a software to reduce the need for human intervention to resolve multi-party conflicts on online social networks. By the means of this software, the auctions in Clarke-Tax mechanism are conducted by the agents representing the users. Since the tax is applied for the bids changing the outcome of an auction, the rational and dominant strategy for an agent is to bid on the actual preference of the user that it represents.

Ulusoy and Yolum [13] propose Privacy Auctioning for Online Social Networks (PANO) where further restrictions are applied on agent-based Clarke-Tax auction mechanism to prevent any manipulation of the system including group-wise handling of budgets and bidding in a limited range. In their follow-up study [14], the auction mechanism is supported by learning agents that can assess the effectiveness of a bid range in the auctions. Further details for this study are provided in the following subsection.

2.2. PANOLA

As the auction mechanisms by themselves fail to guarantee the long term satisfactions of the users especially considering different privacy personas defined over users’ knowledge and motivations [25], Onuralp and Yolum propose PANOLA method [14] which outperforms the underlying PANO auctions [13] realized by agents each of which represents a person, in terms of user satisfactions over time.
In their method, the agents learn the effectiveness of a bid range from both user inputs and the results of previous auctions. They also use two values to estimate the bid effectiveness for a user, namely Value of Content ($V_{Ct}$) and Value of Conformism ($V_{Cf}$). Where the first one is used to determine how important the content is for the user and the other reflects how willing the user is to conform with the group. Based on the calculated effectiveness and user feedback, the utility of a bid range is evaluated which functions as a reward mechanism in the approach.

Although the method is more successful than PANO to satisfy different privacy personas in the group, it does not use the previous bids of other agents in the system. Thus, behaviors of other people in the group are ignored in the calculation of utilities. Furthermore, the method assumes only one type of contents as the subject of the auction, yet this is hardly the case in real-life scenarios. Considering that even motivations of the users that lead to the values mentioned above can vary for different types of the co-owned contents, adding this aspect to the system can make it more robust than the setting which assumes only one type of contents.

2.3. RL under Partial Observability

RL refers to a family of algorithms mostly built on the assumption that the current state that the agent reside in depends solely on the previous one and the action that is taken by the agent, called the Markov property. Markov Decision Process (MDP) formulates the problem as a tuple $\langle S, A, T, R \rangle$, where $S$ is the finite set of states, $A$ is the finite set of possible actions, $T : S \times A \times S \rightarrow [0, 1]$ is the transition function that gives the probability of making transition from a state to another by taking a specific action, and $R : S \times A \rightarrow \mathbb{R}$ is the reward function that gives the reward taken by the agent following some transition. A policy for an MDP is defined as, $\pi : S \times A \rightarrow [0, 1]$, which defines the probability of taking an action in $A$ from a certain state in $S$. The aim of the learning is to find the optimal policy that maximizes the expected cumulative discounted reward of the agent.

If the agent is not capable of capturing all features of the state spaces by having a limited sensory data, then MDP cannot provide a useful formulation for the problem. That’s why a generalization of MDP, Partially Observable MDP is proposed [26]. In this extended formulation, $\Omega$ is introduced as the set of observations that the agent can get within its own capability and $O : S \times A \rightarrow \Pi(\Omega)$ is used as the observation function which gives a probability distribution over possible observations for each action and resulting state.

This POMDP formulation can be interpreted in two different ways. In the first, the assumption is that the underlying MDP is known to the agent, yet it is not able to distinguish its current state. This interpretation leads to the method called belief state [26]. The agent is presumed to be completely unaware of the state space and transition function in the second interpretation. The latter is more realistic for many partially observable problems including ours.

2.4. Multi-Agent Learning

Multi-agent systems have been studied for a long time, especially along with the increase of artificial intelligence applications on the distributed systems [27, 28]. Pioneering studies over multi-agent reinforcement learning (MARL) date back to the early 90s [29]. It
is common to formalize multi-agent reinforcement learning with the Markov game approach following the Littman’s work [30]. This approach is discussed under two subheadings: cooperative and competitive [30,31]. Realization of the multi-agent reinforcement learning settings includes centralized, fully decentralized, and networked decentralized methods [29,32,33]. Each method has its own limitations and theoretical limits [34], yet networked decentralized methods seem to be suitable for the setting we have in mind.

3. Multi-Agent Reinforcement Learning for The Resolution of Privacy Conflicts

In order to build an RL based framework for multi-party conflicts, we follow an incremental approach. Onuralp and Yolum’s PANOLA method [14] is selected as base method in the process. The first steps will be taken to enhance the learning performance by including opponent modeling and different types of contents that are subject to privacy decisions. Within these steps the factors leading to the user preferences will be determined to assess the user utilities better in the auctions. Then these factors will be considered as the main features of state space for the RL problem at hand so that a pure learning solution can be constructed without sacrificing any time or effort for an auction.

3.1. Considering Previous Bids of Others

Previous auctions are included in the learning process of PANOLA, but only by considering them with their results in the agent’s point of view. The bids and privacy choices of others are ignored, and therefore, effectiveness of bids are not affected by these in the calculations. However, if an agent makes use of the knowledge about the support coming from others in the group, it can evaluate this effectiveness better. For instance, if an agent is supported by the majority of the group, then it can use its budget more efficiently by bidding a lower amount.

In the opposite case, if the agent knows that it is not going to be supported by others at all, and let us assume that the content is highly important for its user, then by bidding a higher amount, the agent can get an outcome that is in favor of the represented user. Hence, opponent modeling in this sense can lead a better assessment of user utilities. Furthermore, it can ease the transition of the system to the multi-agent setting as well.

3.2. Introducing Types of Content

Another characteristic of PANOLA creating a possibility for further improvement is that it assumes a single type of content in the auctions. Since even the motivations of the users to provide more inputs to the system can differ from one type to another, considering these types can make the system more applicable to different scenarios in real life.

In its simplest interpretation, the personal values that are used in PANOLA ($V_{CI}$ and $V_{CF}$) can be calculated independently for each type of content. Furthermore, the factors that lead to the user preferences can be introduced into the system at this point to assess these types more thoroughly. At the next phase of the framework these factors can provide sufficient information to the system to decide over the privacy of a content without needing an auction.
3.3. Elimination of The Auction Mechanism

In this subsection we will introduce some of the factors behind the user preferences that can be the main features of the state problem. Since without a proper state representation, any RL solution is bound to fail, these factors are crucial in the elimination of auction mechanism in the system. Hence, the factors must be comprehensive as possible.

For the time being, we categorize these factors into following three groups:

- **Context dependent factors** define the features that depend on the context of the content. For instance, if the user finds sharing a photo taken on their vacation embarrassing then they will want to keep it private. Another user may directly refuse to share a photo based on its quality etc. [35].

- **Time dependent factors** are about timing of the sharing action and/or the creation date of the content. For example, a user may avoid interactions on social media during weekdays. Another person may be unwilling to disclose their whereabouts by sharing a content indicating it for a certain time.

- **Social dependent factors** describe the aspect of the relations between the co-owners and their connections as the audience of the content. As an instance, a user may want to share a photo taken with their close friend, but if a third person who is not so close is also seen in the picture, then the decision may change. Also, these factors include the features that affect the repeating cases of decision making with the same co-owners. For instance, after a user has seen that their friend does not want to share a certain content, they may not prefer sharing a similar content in a future occasion. Hence, these factors lead to the user behaviors which can only be determined by observing the interactions among them.

These factors in any category can be both personal and non-personal. If a factor is personal to a user then most likely it will not be observable by the other agents. However, if a factor is non-personal then it will probably be within the knowledge of all agents. As the factors build the state space of the problem environment, the interactions between each user and representing agent along with the interactions among all agents lead to the observation space. Hence, as long as these interactions are clear, the overall ambiguity in the system will not be preventing an optimal or a near optimal solution for a privacy conflict. By constructing a reward function based on balanced utility along with possibly direct feedback from the users, a long term satisfaction can be achieved by the application of a MARL method over such an observation space. Figure 1 illustrates the proposed environment of MARL setting to resolve MPCs.

The performance of the system can be evaluated by comparing it with PANOLA. Although the most critical metric will be the satisfactions of all users in the long term, the time that is spent for a decision making will also be measured when the realization of the framework is done.

4. Conclusion

In this study, a multi-agent RL setting is presented as a promising candidate to assist the users in solving multi-party privacy conflicts. The setting will be considered as successful, if a long term satisfaction of all users is achieved. Then it will eliminate the need for an auction mechanism in which such a satisfaction is not guaranteed at all by itself.
One can notice that the success of the system that aims to model to users directly depends on their behaviors which are proven to differ what they declare [36,37]. Another limitation is about the comprehensiveness of the features which reflect the user preferences and hence, construct the state space of the problem. Thus, further analysis on these features can lead a system that performs better.

Although the challenging nature of the problem leaves some open questions about the realization of the system, any progress in modeling the users and interacting with them under partial observability will bring benefits to the field beyond the privacy context.

Acknowledgment

This research is financially supported by the Scientific and Technological Research Council of Turkey, https://www.tubitak.gov.tr, through BIDEB 2219 International Postdoctoral Research Scholarship Program; and by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, grant number 024.004.022.

References


Exploring Large Language Models Capabilities to Explain Decision Trees

Paulo Bruno SERAFIM a,1, Pierluigi CRESCENZI a, Gizem GEZICI b, Eleonora CAPPUCCHIO c, d, Salvatore RINZIVILLO d and Fosca GIANNOTTI b

a Gran Sasso Science Institute
b Scuola Normale Superiore di Pisa
c Università di Pisa, Università di Bari Aldo Moro
d Istituto di Scienza e Tecnologie dell’Informazione, CNR

Abstract. Decision trees are widely adopted in Machine Learning tasks due to their operation simplicity and interpretability aspects. However, following the decision process path taken by trees can be difficult in a complex scenario or in a case where a user has no familiarity with them. Prior research showed that converting outcomes to natural language is an accessible way to facilitate understanding for non-expert users in several tasks. More recently, there has been a growing effort to use Large Language Models (LLMs) as a tool for providing natural language texts. In this paper, we examine the proficiency of LLMs to explain decision tree predictions in simple terms through the generation of natural language explanations. By exploring different textual representations and prompt engineering strategies, we identify capabilities that strengthen LLMs as a competent explainer as well as highlight potential challenges and limitations, opening further research possibilities on natural language explanations for decision trees.

Keywords. Explainable AI, decision tree, natural language generation

1. Introduction

The high complexity of novel Machine Learning (ML) models is an important factor in the production of remarkable results since simpler models were not able to achieve the same level of performance. However, as the models grow, there have been growing concerns due to the opaque nature of widely used black-box algorithms. Among the problems that arise from the non-transparency are the trustworthiness, fairness, and accountability of the models [1]. In this context, the need for an approach to bridge the gap between the algorithmic perspective and human-centered explanation is evident.

Among key explanatory elements from disciplines such as social sciences and psychology, interactivity has been described as a core value for the implementation of human-centric explainable AI applications [2]. The explanation process has to be outlined as a continuous dialogue between a sender and a receiver, making the user engage in a conversation [3,4]. More recently, Large Language Models (LLM) have shown an exceptional ability to generate natural language responses [5]. As such, LLMs emerge

1Corresponding Author: Paulo Bruno Serafim, paulo.desousa@gssi.it.
as a promising tool to maintain an active dialogue with a user whose goal is to better understand a model decision process.

Decision trees are widely used in classification tasks and recognized for their simplicity, being considered a method interpretable by nature [6,7]. Nonetheless, even with decision trees with a large number of nodes or if the user is not familiar with trees [8]. Both situations demonstrate the usefulness of having a tool to summarize an explanation of the decision taken in non-technical terms. In this paper, we explore the capabilities of LLMs to explain decision trees for non-expert users.

Our contributions are as follows. First, we expand on prior work of generating Natural Language Explanations for decision trees using LLMs by proposing a general explanation pipeline. We explore prompt engineering strategies to evaluate whether the generated explanations are coherent, informative, and easy to understand. We also evaluate how sensible is the LLM to different text representations. Finally, we present further research directions that go toward more human-centered explainable artificial intelligence.

2. Related Works

Generating Natural Language Explanations (NLE) for Machine Learning models has been an active research topic [9]. Among works that generate NLE, we highlight the ones that use decision trees as the classifier. López-Trigo et al. [10] developed a pattern-matching method to generate local and global explanations following specific grammar rules. As an extension of the previous work, Alonso et al. [11] developed the ExpliClas web application to provide explanations to the users.

Graviidilis et al. [12] reported a method to generate natural language explanations by using a Natural Language Generation component in the pipeline. The tree structure is used to define a plan of action, which is then converted to a set of contextualized rules. More recently, Ziems et al. [13] used an LLM to generate NLEs for classification from decision trees trained in a Network Intrusion Detection (NID) problem. The dataset description, feature descriptions, and textual representation of the decision tree are combined to form an input text which is passed to an LLM to provide explanations.

In this paper, we present a method to generate NLE from decision trees using LLMs for any task. We extend the work of Ziems et al. [13] by generalizing the pipeline to further problems, without specific requirements of NID. Moreover, we do not want the explanation to provide background information or technical answers, nor do we ask for them. Compared to the mentioned works, our goal is to explore LLM capabilities to obtain explanations that are easier to understand by a non-expert user.

3. Methodology

There are several steps involved in the process of producing an appropriate natural language explanation from an LLM. First, it is important to note that we are not using models with multi-modal capabilities, as such we only interact with the LLM through textual inputs. In order to obtain adequate responses, it is necessary to provide a well-structured prompt that contains all the required information. We construct the input following a series of subprocesses that leverage values from the task and the instance, supplemented by information that can improve explanation quality.
In the general framework, we train a decision tree from the task dataset and then generate a textual tree representation. Together with task features, classes, and an instance, we pass them as a prompt to an LLM. This approach is similar to the one present in Ziems et al. [13], but it is not restricted to Network Intrusion Detection as it was extended to generate explanations from decision trees trained on any task. Lastly, the LLM outputs an explanation, which concludes the generation pipeline, as illustrated in Figure 1.

The prompt contains all the relevant information for the LLM to generate the explanation in a textual format. We predefined a prompt pattern to construct the complete input text by adding specific information from the dataset, tree, and instance. Additionally, we include a demonstration and instructions to guide the LLM in generating the explanation. Each prompt component is described in detail in the following sections and a complete example can be found in Table 1.

**Dataset Features and Classes.** At the beginning of the prompt, we list the dataset feature names, which are directly extracted from the dataset. The class names comprise the second input to the prompt pattern. Each dataset may have a different number of classes and similar to the feature names, class names are not necessarily available. Together with the features, they form the dataset description.

**Decision Tree Representation.** As a graph, there are many ways to represent a decision tree. The most common is to use a diagram formed from nodes and directional edges. However, since we are using textual inputs to the LLM, it is necessary to transform the tree into a textual representation. Since we do not predefine a tree template, any externally generated textual representation of a decision tree can be used.

**Demonstration.** Prior work has shown that LLMs tend to give more appropriate responses when a demonstration is passed in the prompt [14]. As such, we follow these suggestions for better prompting and include an example to be used by the LLM as a guideline for its answer. We created one sample by ourselves which is a depiction of an expected explanation from a given instance.

**Instance Description.** The instance textual representation is formed by concatenating all feature names and their corresponding values, which is automatically done by the prompt construction script. Therefore, all feature names match exactly the ones that are present in the dataset description. Note that although the instance values have to be consistent with the features, it is not required to be a real instance from the dataset.

**Instructions.** Once again we refer to good prompt engineering strategies [14,5] in order to obtain more suitable explanations by giving clearer instructions to the LLM. Similar to Ziems et al. [13], we ask the LLM to avoid mentioning the mechanics of the tree and answer in natural language. Additionally, we expect all relevant features to be mentioned, while ignoring unused ones. Finally, we include a sentence to prevent numbers and favor more general terms like ‘high’ and ‘low’ instead.
Question. The last part of the prompt is the question itself. At this moment, all descriptions were extracted and additional information was assembled. Then, we finally ask the LLM to generate explanations that are similar to the demonstration while following the given instructions. An example of a complete prompt can be found in Table 1.

Table 1. Example of prompt constructed according to the pattern.

Consider a dataset that has the following features: sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm). Each instance can be classified into one of the following classes: setosa, versicolor, and virginica. A decision tree was trained on the dataset and the following tree was obtained:

```
|--- petal length (cm) <= 2.45
  |--- class: 0
  |--- petal length (cm) > 2.45
    |--- petal width (cm) <= 1.55
      |--- petal length (cm) <= 4.95
        |--- class: 1
        |--- petal length (cm) > 4.95
          |--- class: 2
          |--- petal width (cm) > 1.55
            |--- petal width (cm) <= 1.70
              |--- class: 1
              |--- petal width (cm) > 1.70
                |--- class: 2
```

Given an instance of the iris dataset with features: sepal length (cm) = 3.6, sepal width (cm) = 0.8, petal length (cm) = 4.3, and petal width (cm) = 2.9, and a confidence value of 57.81%, a good explanation for why the instance was classified as virginica is: ‘By evaluating the feature values, it is possible to observe that both petal length and petal width are high. This means that by following the decision tree path, the instance should be classified as virginica, although the tree is not very confident in this result.’

An instance has features: sepal length (cm) = 7.3, sepal width (cm) = 2.9, petal length (cm) = 6.3, and petal width (cm) = 1.8.

To answer the following question, do not refer to the underlying mechanics of the decision tree in any way, and only refer to the features using natural language. All the relevant features must be mentioned in the answer, but features that were not used by the tree should be ignored. Moreover, do not use any technical jargon or numerical values in the response and prefer to use terms like ‘high’ and ‘low’.

Please explain in similar terms why the decision tree concluded that the given example is virginica with a confidence of 97.06%.

4. Experiments

In this section, we show that LLMs are able to generate appropriate explanations in natural language from decision trees. However, to obtain good explanations, the prompt must be carefully constructed, since LLMs are known to be sensitive to the input question. We performed experiments to evaluate different prompt structures until achieving the one that is presented in Section 3. We also evaluate LLM sensibility to different textual representations of the prompt components and their influence on the explanation quality.

4.1. Experiment Settings

Following the process described in Section 3, the first step is to train a decision tree given a task dataset. We used the scikit-learn toolkit [15] to train a decision tree and obtain information about feature names, class names, and an instance. The experiments have been performed to evaluate different classification datasets, although we only report results on “iris” for succinctness. It is important to note that in the scope of this paper, our focus is not on model performance but rather on the decision tree structure. Any externally trained decision tree can be used and the same process can be applied to any classification dataset as long as its textual representation is passed to the prompt construction.
Since our goal is to evaluate the capabilities of LLMs and not to compare them, we opted to focus on a single model and observe the quality of the results we can achieve. We used “gpt-3.5-turbo” in all experiments, which is a variant of the GPT-3 model [14] and currently used in the ChatGPT web application [16], through the OpenAI API. We did not perform any fine-tuning process or further modifications to the pre-trained model. No system instruction was used and all the prompt was passed as a user request. Therefore, we expect that similar results can be obtained by users without API access.

4.2. Prompt Engineering

To generate good explanations, we tested different prompt engineering strategies. Our goal was to evaluate which kind of prompt would give more adequate results.

**Direct Question.** The most basic way of interacting with an LLM is by simply asking a question. Our baseline approach follows this principle, in which we directly ask the model to generate a natural language explanation. In this experiment, only the dataset description, the decision tree textual representation, the instance, and the question have been passed to the prompt.

The model response: “The decision tree concluded that the given example is virginica with a confidence of 97.06% because it followed a series of rules based on the features of the instance. First, it checked if the petal length was less than or equal to 2.45 cm, which was not the case for the given example. Then, it checked if the petal width was less than or equal to 1.55 cm, which was also not the case. Finally, it checked if the petal width was greater than 1.70 cm, which was true for the given example. Based on these rules, the decision tree determined that the given example belongs to the class virginica with a high confidence of 97.06%.”

The answer follows a sort of rule-based approach, listing for each decision node why a path was taken. This example shows that simply asking for an explanation is not sufficient to obtain a response that better helps non-expert users in comparison to the decision tree graph representation, although the reasoning was correct.

**Demonstration.** A well-known strategy to improve the output quality is to provide demonstrations since LLMs have been proven to be able to generate outputs that are similar to given examples [14,5]. The second experiment is based on this approach. In addition to the previous prompt components, we provide a demonstration, as shown in Table 1, and ask the LLM to generate a “similar” explanation instead.

The model response: “The decision tree classified the given example as virginica with a confidence of 97.06% because it evaluated the feature values and found that both petal length and petal width are high. According to the decision tree path, when petal length is greater than 2.45 and petal width is greater than 1.55, the instance is classified as virginica. In this case, the high values of petal length (6.3 cm) and petal width (1.8 cm) led the decision tree to confidently classify the instance as virginica.”

The output clearly shows an explanation more suited to non-expert users, with fewer technical terms. For example, the first sentence presents a summary of the classification mentioning “high” feature values. We believe that this new explanation, which does not contain decision rules, would be more comprehensible for a user who has no familiarity with decision trees. Nonetheless, there are still numerical values present, and including the tree path adds unnecessary complexity to the answer.

**Provide Instructions.** Finally, following the idea of making it clear to the LLM what we want to achieve, we provide instructions after the instance description and before
the question. Our goal is to make clear in the prompt what should be included in the answer and what should be avoided. As a result, we obtained the following response from the model: “The decision tree concluded that the given example is virginica with a confidence of 97.06% because the petal length is high and the petal width is also high. This combination of high petal length and petal width led the decision tree to classify the instance as virginica with high confidence.”

Without mentioning the path taken, decision rules, and feature values, the current response is easier to understand. It is also clear and succinct, which makes it more suitable for non-expert users. Thus, we proceed with the experiments by modifying the text while maintaining the prompt structure.

4.3. Textual Characterization

In this section, our goal is to evaluate the LLM sensitivity to distinct textual representations of the prompt components. More specifically, we analyze how much the explanations change if we keep the same prompt structure, but modify the text input.

Incorrect Classification. The first question we would like to answer is: does the LLM recognize that a given instance was incorrectly classified? To answer this question, we change the instance prediction in the prompt to an incorrect one. We keep the same feature values shown in Table 1, but we ask for an explanation for the instance being classified as versicolor. The question is: “Please explain in similar terms why the decision tree concluded that the given example is versicolor with a confidence of 97.06 %.”

The answer obtained was: “The decision tree concluded that the given example is versicolor with a confidence of 97.06% because the petal length is high and the petal width is low. This combination of feature values led the tree to classify the instance as versicolor with high confidence.”

The LLM does not recognize that the instance was incorrectly classified, which indicates that the LLM is not checking the correctness of the prediction. More interestingly, the LLM changed its judgment about one feature so that it could match the given prediction, as shown in the part “petal width is low” which has been changed to “petal width is high”. Note that if the petal width was lower than 1.55, the instance would be classified as versicolor, which is the class that matches the given prediction. In a sense, the LLM changed its internal generation tendencies to agree with the given prediction.

Tree Representation. In the next experiment, we return to use the correct classification, but we change the tree representation. Instead of using the textual representation given by scikit-learn, we generate a different representation of the same tree using the ChatGPT web application [16].

We obtained the answer: “The decision tree classified the given example as virginica with a high confidence because the petal length is high and the petal width is also high. This combination of high petal length and petal width led the decision tree to conclude that the instance belongs to the virginica class.” As we can see, the response is almost identical to the one using the previous tree text. This indicates that by using the current prompt pattern, the generated explanation is agnostic to the tree representation.

External Explanation. In order to evaluate how the demonstration affects the explanation, we alter the example provided. We replace the previous text with an example taken from López-Trigo et al. [10]. Keeping the rest of the prompt the same, we obtain the following answer: “The decision tree concluded that the given example is virginica
with a confidence of 97.06% because its petal width is high, which is a key factor in determining the class. Additionally, its petal length is high, which also contributes to the classification as virginica. These features align closely with the criteria used by the decision tree to classify instances as virginica.” As we can see, although there are minor differences, the generated explanation is still very similar to the previous one.

By observing that the instructions are an important factor, we decide to modify them and evaluate if the explanation changes. Therefore, we removed the last sentence of the instructions and kept the rest of the prompt the same. The answer obtained was: “The decision tree concluded that the given example is virginica with a confidence of 97.06% because its petal width (1.8 cm) is greater than the split value (1.7 cm), and its petal length (6.3 cm) is also greater than the split value (4.9 cm). These features align with the criteria used by the decision tree to classify instances as virginica.”

The output clearly shows that the LLM tried to follow a similar pattern as the given example, using terms like “split values” as well as showing the values inside parentheses. This result indicates that although demonstrations can be used to guide the LLM, the instructions are the most influential part of the prompt.

**Demonstration for Different Task.** In the last experiment, we analyze the generalization capabilities of the LLM. More specifically, we verify if the LLM is able to generate a natural language explanation given a demonstration from a different task. We change the task to explain the classification of an instance from the “wine” dataset. However, we still give the LLM a demonstration from the iris dataset. This is an example of obtained output: “The decision tree classified the given example as class 0 with 100.00% confidence because the value of color intensity is low, and the value of ash is high. This led the tree to directly classify the example as class 0 without considering any other features.”

As we can see, the LLM is able to generate a coherent explanation for the classification of the instance from the wine dataset. This is a very interesting result, as it shows that the LLM can generalize the explanation generation process to different tasks while benefiting from the demonstration given in the prompt.

## 5. Discussion

After performing several experiments, the results show that Large Language Models are able to explain decision trees in a way that can be understood by non-expert users. It is also clear that the structure and given instructions are more influential than the textual representations or the demonstration. Below we discuss some limitations and positive aspects from the results we obtained, as well as envision some future research directions.

### 5.1. On the Capabilities and Limitations of LLM Explainers

By following the pipeline described in Section 3, we were able to obtain good explanations from decision trees. Both the prompt structure and the given instructions were influential in the quality of the responses. The best explanations clearly stated the reasons for the classification without using feature values or resorting to the tree path. The collected evidence shows that LLMs can be good explanators but they present some limitations.

For example, the responses might contain technical terms or hard-to-follow sentences, unless the LLM is directly instructed not to do so. However, if we anticipate some
desired explanation, the LLM might generate appropriate outputs. Another problem is that the generated explanations are not grounded in the truth, which makes the model provide incorrect predictions. Since the LLM was not able to check correctness, it will change the answer to agree with the user, even if the path is not followed correctly.

Nevertheless, one experiment showed that the output was agnostic to the textual representations of the tree, indicating the robustness of the method. Moreover, the LLM was also able to generalize an explanation given an example from a different task. In short, based on our experiments, LLMs show capabilities that allow them to be used as explainers, such as generating text adequate for non-expert users. However, they also present several limitations, including providing incorrect explanations, which constitutes a risk to use them in sensible circumstances.

5.2. Future Research Directions

The results obtained in this work open several research directions. A natural follow-up is to better evaluate the capabilities of LLMs by performing more robust experiments. For example, the structuring of inputs and demonstrations can be compared with recent prompt engineering strategies. Another clear continuation is to expand the framework to other LLMs, such as GPT-4 [17], Llama [18,19], and Gemini [20]. This would allow us to evaluate the quality of the explanations generated by different models.

Another interesting direction is to apply the framework to other classification models, such as KNNs, SVMs, and Bayesian networks. However, in order to evaluate the results among them, it is necessary to have a formal definition of the explanation quality with the corresponding metrics to automatically perform this analysis. Although not straightforward, it would allow the comparison of completely different models and tasks.

Prior work has shown the generation of explanations for black-box models using surrogate decision trees, such as LORE [21]. Since the generation pipeline presented in this paper shows promising results in small trees, it would be interesting to evaluate how it performs in larger trees, which are more common in typical cases. This would enable the generation of natural language explanations for a wide range of tasks and the ability to provide information that is returned from local explanations, like counterfactuals [22].

Finally, the goal of working in generating natural language explanations is to make the models more transparent and understandable to non-expert users. Therefore, the next step would be to enable a bi-directional interaction channel between the user and the LLM. This would allow the user to ask questions and clarify existing doubts, and the LLM to ask for more information when the prompt is not clear enough. To achieve this goal, it is also necessary to couple the generation methods with user evaluation studies. Research on Human-Centered XAI [4,23] indicates that this is a promising direction.

Acknowledgements

This work is supported by the European Community under the Horizon 2020 programmes: ERC-2018-ADG G.A. 834756 XAI, by HumanE-AI-Net GA 952026, and the NextGenerationEU programme under the funding schemes PNRR-PE-AI (M4C2, investment 1.3, line on AI) FAIR (Future Artificial Intelligence Research) Prot. PE0000013. Research partially funded by project GAMING “Graph Algorithms and MiniNg for Green agents”, National Recovery and Resilience Plan, NextGenerationEU. We would like to thank Noah Ziems and his co-authors for sharing their code.
References


An Interactive Interface for Feature Space Navigation

Eleonora CAPPUCIO\textsuperscript{a,b,c,1}, Isacco BERETTA\textsuperscript{a}, Marta MARCHIORI MANERBA\textsuperscript{a,c} and Salvatore RINZIVILLO\textsuperscript{c}

\textsuperscript{a} Università di Pisa
\textsuperscript{b} Università degli Studi di Bari Aldo Moro
\textsuperscript{c} CNR-ISTI

ORCiD ID: Eleonora Cappuccio https://orcid.org/0000-0002-6105-2512, Isacco Beretta https://orcid.org/0000-0002-0463-6810, Marta Marchiori Manerba https://orcid.org/0000-0003-2251-1824, Salvatore Rinzivillo https://orcid.org/0000-0003-4404-4147

Abstract. In this paper, we present Feature Space Navigator, an interactive interface that allows an exploration of the decision boundary of a model. The proposal aims to overcome the limitations of the techno-solutionist approach to explanations based on factual and counterfactual generation, reaffirming interactivity as a core value in designing the conversation between the model and the user. Starting from an instance, users can explore the feature space by selectively modifying the original instance, on the basis of her own knowledge and experience. The interface visually displays how model predictions react in response to the adjustments introduced by the users, letting them to identify relevant prototypes and counterfactuals. Our proposal leverages the autonomy and control of the users that can explore the behavior of the decision model accordingly with their own knowledge base, reducing the need for a dedicated explanation algorithm.

Keywords. Human-Centered AI, Human-AI Interaction, Interactive ML, Explainable AI, Counterfactuals, Explanation User Interface

1. Introduction

Progress in the performance and efficiency of automatic decision-making systems has incentivized AI-based solutions in pervasive and impactful contexts of daily life, such as finance, healthcare, and transportation. A problematic aspect of these models is the lack of explainability: it is difficult to provide the reasons behind the automated decisions due to the complexity of the algorithms and the large amounts of data processed. This lack of transparency and accountability can lead to mistrust and concerns about the ethical implications of using AI in decision-making [1]. In some sensitive real-world contexts, accounting for the algorithmic decision is necessary for the user to understand and contest the motivations behind the system’s output. This is especially important when the outcome strongly impacts human life or has harmful consequences. Moreover, relevant

\footnote{1Corresponding Author: Eleonora Cappuccio, eleonora.cappuccio@phd.unipi.it.}
policies such as the GDPR (General Data Protection Regulations) [2,3] claim the relevance of appealing in case of rejection, i.e., to request what changes users should make to be accepted in receiving a positive response, for example, to a loan application. In this way, users can be empowered to understand how the algorithmic decision affects them.

In this work, we present Feature Space Navigator, an interactive interface that allows an exploration of the decision boundary of any black box model. We report in Fig. 1 a visual workflow of our approach. Imagine a person applying for a loan to a bank that leverages an automatic decision-making system. The instance $x$ represents her current situation (stage [1] of the workflow): she has a certain amount of capital, her current working position is being a chef in a restaurant, and she has no degree. The decision emitted by the black box model is negative, hence her request is denied.

Following the diagram, in stage [2a] an automatic counterfactual technique generates instances trying to optimize a predefined cost function that embeds a set of constraints, e.g., plausibility and diversity. Three counterfactual instances receiving a positive outcome are proposed to the user (stage [3a]). The scenarios suggested are: (i) the person has a medium amount of capital and occupies a business role, (ii) the person remains a chef, but she has a high amount of capital, and (iii) the person has a medium amount of capital, she is a doctor and she has a degree in medicine. From the counterfactuals returned, the user satisfaction is symbolized by the smiley and gloomy faces.

Following this pipeline, the subjective human satisfaction is impossible to capture and it remains unknown or an irrelevant feedback.

Our proposal, exemplified in stage [2b], consists of an interface that lets the users freely explore their options, changing the features at will. The set of constrains is not directly embedded nor expressed through the interface, but it is implicitly delegated to the common sense of the users. The resulting counterfactual user-crafted in stage [3b] has the same amount of money, by profession she is a florist and has a degree. The difference between our approach w.r.t. the traditional one consists in the possibility for the user to...
The rest of the paper is organized as follows. In Section 2, we report a critical literature discussion and a review of the state-of-the-art solutions that closely align with our contribution. Section 4 presents the proposed method. In Section 5, we report several examples testing different datasets to demonstrate our proposal’s novelty and effectiveness for counterfactual generation. Finally, Section 6 indicates future research directions and discusses the limitations of our approach.

2. Related Work

2.1. Counterfactual generation techniques

Counterfactual approaches explain individual predictions describing what-if contrastive scenarios [4,5,6]. Specifically, the explanations indicate to the user the minimum change necessary in the feature space representing them so that the output of the automatic system changes toward the desired outcome.

Various approaches have been devised for generating counterfactuals, including optimization, heuristic search, instance-based, and decision trees. We refer to [7] for a more comprehensive overview. Evaluating the quality of counterfactual explanations requires careful consideration of multiple desired qualities, which have been defined in the literature as validity, sparsity, similarity, plausibility, discriminative power, actionability, causality, and diversity [8,9,10]. The complexity of this range of evaluation metrics presents a critical challenge in comparing the currently available algorithms and establishing benchmark procedures. Moreover, most of the current counterfactual generation techniques do not allow for user interaction, thus limiting the practical applicability of these methods in real-world contexts [11].

2.2. Human-Centered Approach

It is crucial to rethink new ways of interaction between humans and the XAI algorithm, focusing more on the user’s point of view and needs [12,13,14,15]. The effectiveness of an explanation is based on the extent to which the recipient regards it as comprehensible and useful. A human-centered approach is needed to fill the gap between the algorithmic perspective and an explanation that is truly intelligible to humans [13]. In fact, as stated Liao et al. [13], XAI presents more of a design challenge than an algorithmic challenge.

Previous social science research has shown that human explanations are often contrastive. Therefore, counterfactuals are generally considered less cognitively demanding and more intuitive [16]. Given these assumptions, our proposed interface and algorithm are outlined as starting points for several case studies, see Section 5, and are therefore designed for domain experts and lay users [17].

In 2018 the DARPA explainable AI initiative framed the explainable AI process as a three-stage approach, distinguishing between the explainable model, the explanation user interface. [18]. This definition draws a difference between the model through which an explanation of a machine learning algorithm is generated and the means used to communicate it to the user [19]. From this perspective, the explanation process has
to be outlined as a continuous dialogue between a sender and a receiver [14]. Therefore XAI has to consider interactivity as a fundamental part of the process [20,12,19], for example, through the design of novel interfaces that allow model inspection at will [21], and promote an understanding of the computational processes giving users control of the algorithmic actions [22].

2.3. Interactive explanations and visualizations

In the following, we frame Feature Space Navigator in relation to existing literature solutions, and critically discuss the novelty and positioning of our contribution.

DECE is an advanced counterfactual generation method tailored for both global and local interpretability tasks [23]. With a strong focus on visualization and user interface design, DECE automates counterfactual generation using established metrics such as diversity and validity. Notably, it facilitates user-driven constraint imposition through a decision tree framework, offering fine-grained control over the generation process. Moreover, DECE extends beyond dataset limitations by exploring counterfactuals independent of available data, while leveraging subset counterfactuals to dissect machine learning model decision boundaries.

Prospector is a method designed to generate instance counterfactuals via a visual interface [24]. It utilizes a semi-global score, specifically a partial dependence plot, to estimate the probable prediction when specific features of the original instance are modified. This process involves three distinct phases: patient selection, patient inspection, and the generation of partial dependence plots. Access to both the model and the dataset used for training is required. The tool is specifically designed for data scientists. It allows them to understand how features affect the prediction, thus improving a predictive model.

DiCE framework generates and evaluates counterfactuals according to two properties, namely plausibility and diversity [25]. The authors test DiCE on four real-world datasets and compare its counterfactuals with popular local explanation methods and other prior approaches.

T-LACE generates reliable counterfactual explanations leveraging on two properties of a tailored transparent latent space, namely similarity and linearity [26]. An interactive framework for auto-encoders allows a visual exploration of the latent space, shedding insights on the links between input features and model prediction [27].

For a thorough overview, we refer to [28], where authors present techniques and strategy designed to explore and explain model’s predictions at instance-level, i.e., for a single record, and at dataset-level, i.e., for an entire collection.

Our contribution We outline some of the distinguishing properties of Feature Space Navigator, our method, setting it apart from existing literature. Firstly, it prioritizes the user, who not only operates the tool but also indirectly defines the optimization metric, ensuring subjective considerations are captured. Moreover, it requires minimal technical skills, as no coding knowledge is needed, and users interact through intuitive sliders. It works with any model and does not require access to the dataset, making it easy to explore different models without constraints. Furthermore, it is purely local, focusing solely on the specific instance at hand, ensuring relevance to the user’s context. Users can adjust instances feature by feature, enabling thorough exploration of the feature space. Overall, our method is user-friendly, adaptable, and effective for exploring machine learning systems.
3. Setting the Stage

**Preliminaries**  Given a classifier $b$ that outputs the decision $y = b(x)$ for an instance $x$, a counterfactual explainer $E$ outputs a perturbed instance $x'$ such that the decision of $b$ on $x'$ changes, i.e., $b(x') \neq y$, and such that the cost of the action $c(x, x')$ to go from $x$ to $x'$ is minimal. Minimality refers to an abstract cost function that must be specified when implementing any concrete method. Often, the choice falls on an $\ell_p$ norm [11]. In the vast majority of cases $c$ exhibits the following limitations:

**Limitation 1. Stationarity:** $c(x, x')$ is predetermined and not updated through user feedback. Moreover, it lacks the ability for $E$ to adapt during post-deployment usage, which would allow for fine-tuning between user needs and the tool’s effectiveness (see Research Challenge 9 in [29]).

**Limitation 2. Translational Invariance:** $c(x, x')$ depends solely on the distance between $x$ and $x'$, i.e., $c(x, x') = c(x' - x)$. For example, increasing the salary by 300£ is assumed to be equally difficult for an individual earning 1000£ and one earning 3000£.

**Limitation 3. Universality:** $c$ cannot depend on exogenous factors of the system. In other words, it is assumed that the same cost function is suitable for all users: in reality, individual properties not visible to the system can influence $c$. For example, the ease of changing jobs may depend on the types of available activities in the residential area. The cost of an action in the real world depends on individuals and multiple, often subjective, factors. Ignoring this aspect often proves to be overly restrictive (see Research Challenge 10 in [29]). An attempt to address this limitation can be found in [30].

Our contribution aims to address these limitations, especially from a user-focused viewpoint.

4. Feature Space Navigator

Building upon [31], our method proposes an alternative formulation of the counterfactual generation problem, departing from the algorithmic perspective and instead shifting the focus to user interaction. Our proposal leverages the autonomy and control of the users that can explore the behavior of the decision model accordingly with their own knowledge base, reducing the need for a dedicated explanation algorithm. From this perspective, users are designated as the “domain experts” of their own real-world scenario.

**Feature Space Navigator** is local, post-hoc and agnostic, i.e., designed to account for any black box. It offers an interactive interface of the areas where the model returns the desired outcome by providing a graphical visualization. The entire process is delegated to the users that are free to explore the feature space of the system and generate autonomously, “by hand”, effective counterfactual and prototype instances that they deem feasible, desirable, and suitable to their specific evaluations and needs. Users are therefore empowered with complete autonomy and control over the process without the mediation of an explanation algorithm between the human and the decision model.

---

2Feature Space Navigator is available at https://github.com/Elecapp/Feature_space_navigator.

3User interactions through the interface are not stored; hence, no concern occurs regarding the safety of private information that users might input.
Algorithm 1: GRID EV ALUATION(ranges, bins, X, x, model)

Input : ranges - features ranges, bins - number of bins, X - dataset, x - original instance, model - black box

Output: \( \hat{Y} \) - model’s predictions

1. \( d \leftarrow \text{len}(x); \quad \text{// storing number of features} \)
2. \( \hat{Y} \leftarrow [0]_{d \times \text{bins}}; \quad \text{// creating a matrix of zeros} \)
3. for \( i = 0 \) to \( d \) do
   4. if \( \text{categorical}(x_i) \) then
      5. for \( v \in \text{unique}(X_i) \) do
         6. \( \hat{x} \leftarrow \text{copy}(x); \quad \text{// creating a copy of x} \)
         7. \( \hat{x}_i \leftarrow v; \quad \text{// changing \( \hat{x} \)'s \( i \)-th feature with the new value} \)
         8. \( \hat{Y}_{i,b} \leftarrow \text{model}(\hat{x}); \quad \text{// predicting score} \)
   9. else
      10. \( \text{min, max} \leftarrow \text{ranges}_i; \quad \text{// saving min and max of current feature} \)
      11. for \( b = 0 \) to \( \text{bins} \) do
         12. \( v \leftarrow \text{min} + \frac{b}{\text{bins}} \times (\text{max} - \text{min}); \quad \text{// computing the new value} \)
         13. \( \hat{x} \leftarrow \text{copy}(x); \quad \text{// creating a copy of x} \)
         14. \( \hat{x}_i \leftarrow v; \quad \text{// changing \( \hat{x} \)'s \( i \)-th feature with the new value} \)
         15. \( \hat{Y}_{i,b} \leftarrow \text{model}(\hat{x}); \quad \text{// predicting score} \)
16. return \( \hat{Y} \)

the following, we outline the interface, the underlying algorithm, and discuss evaluation aspects of the outputs returned by Feature Space Navigator.

Interface The interface, depicted in Section 5, visually displays how model predictions react in response to the adjustments introduced by the users, letting them to identify relevant prototypes and counterfactuals. It allows users to independently explore the feature space by selectively modifying the original instance on the basis of her own knowledge and experience, i.e., changing the value of one of the features that characterize the instance \( x \). This can be done introducing minimal changes through the adjustment of the slider feature by feature within its designated range. Specifically, each slider is characterized by a colored gradient representing the model’s scores for the different values of the feature. Users can choose which feature to modify by moving the cursor to a value that improves their score, selecting the feature they find most comfortable to change. After each adjustment, the instance \( x \) is promptly updated with new values, generating new gradients and enabling further modifications until the user is content with the set of changes and insights. Through a number of attempts, generating both prototypes and counterfactuals, the user intuitively guesses the feature importance of the classifier decision boundary through gradients change.

Algorithm The user can move along each axis in the feature space but can not make diagonal movements. The cumulative nature of the modifications ensures the ability to reach any point in the space. In Algorithm 1, we report the pseudo-code for a single step of the process. It requires access to the model’s probability outputs and is specifically tailored for tabular data with a binary target variable. The number of predictions re-
quired scales linearly with the number of features, ensuring the interface update remains essentially real-time.

5. Scenarios

**Dataset** In this work, we implemented Feature Space Navigator on a dataset widely used in academic research and freely accessed through the UCI Repository [32]: the STATLOG (GERMAN CREDIT DATA) dataset [33], which categorizes individuals as having good or bad credit risk through a binary classification task. It comprises 1000 instances and 20 features, all of which are categorical and integer types. The categorical features are handled in the interface by preserving the original categorical feature’s record alongside the value generated by the LabelEncoder.

**Classifiers** We performed an 85-15 split in train and test for STATLOG and a 70-30 split for WINE. We ran our evaluation on Random Forest implemented by the scikit-learn library4. The best parameters are identified through a Randomized Search over a dictionary of possibilities. Both the classifiers were performed using a fixed random state.

**Understanding the Effects of Interventions** In the following, we provide examples of potential usage of the interface on the adopted datasets, aiming to illustrate the functionality of Feature Space Navigator.

**STATLOG Scenario** Let us first examine a case from the STATLOG dataset, where a user applies for a loan. The initial situation, depicted in Fig. 2a, receives a negative outcome (the credit risk is bad, therefore the loan is not granted). The user, after scrutinizing the interface, decides to lower the credit amount requested, as documented in Figure 2b. Since the outcome issued by the black box remains negative, the user is not satisfied and applies a different change, this time adjusting the duration in month of the loan (Figure 2c). In this case the user finally receives a good credit risk classification.

6. Conclusion

In this work, we have developed Feature Space Navigator, an interactive graphical interface that facilitates exploration of the feature space of the decision boundary of a black box, providing an intuitive visualization of the areas where the model produces the desired outcomes. This user-centric approach enhances transparency and control in the decision-making process, contributing to the advancement of XAI methodologies.

Embracing a human-centered perspective, the evaluation requires a conceptual shift by involving users within it in order to understand whether the explanation is convincing for the intended user, the ultimate stakeholder in the process [34,35]. Concretely, a qualitative evaluation in the form of a user study is necessary. The study will provide a measure of human validation regarding the comprehensibility and usability of the interface, as well as the usefulness of exploration as a means of generating explanations for the underlying decision model.

(a) Starting situation.

(b) User changes Credit Amount; credit risk is still bad.

(c) User changes Duration in Month; credit risk becomes good.

Figure 2. STATLOG Scenario.
Acknowledgements

This work has been supported by the European Community Horizon 2020 programme under the funding scheme ERC-2018-ADG G.A. 834756 XAI: Science and technology for the eXplanation of AI decision making, and the NextGenerationEU programme under the funding schemes PNRR-PE-AI scheme (M4C2, investment 1.3, line on AI) FAIR (Future Artificial Intelligence Research) Prot. PE00000013.

References


Chromik M, Schuessler M. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. Exx@ace@ iui. 2020;54.


[33] Hofmann H. Statlog (German Credit Data); 1994. DOI: https://doi.org/10.24432/C5NC77. UCI Machine Learning Repository.


Human-Centered AI for Dementia Care: Using Reinforcement Learning for Personalized Interventions Support in Eating and Drinking Scenarios

Wen-Tseng CHANG\textsuperscript{a}, Shihan WANG\textsuperscript{b}, Stephanie KRAMER\textsuperscript{c}, Michel OEY\textsuperscript{a}, Somaya BEN ALLOUCH\textsuperscript{a, c, 1}

\textsuperscript{a}Amsterdam University of Applied Sciences
\textsuperscript{b}Utrecht University
\textsuperscript{c}University of Amsterdam

Abstract. For people with early-dementia (PwD), it can be challenging to remember to eat and drink regularly and maintain a healthy independent living. Existing intelligent home technologies primarily focus on activity recognition but lack adaptive support. This research addresses this gap by developing an AI system inspired by the Just-in-Time Adaptive Intervention (JITAI) concept. It adapts to individual behaviors and provides personalized interventions within the home environment, reminding and encouraging PwD to manage their eating and drinking routines. Considering the cognitive impairment of PwD, we design a human-centered AI system based on healthcare theories and caregivers’ insights. It employs reinforcement learning (RL) techniques to deliver personalized interventions. To avoid overwhelming interaction with PwD, we develop an RL-based simulation protocol. This allows us to evaluate different RL algorithms in various simulation scenarios, not only finding the most effective and efficient approach but also validating the robustness of our system before implementation in real-world human experiments. The simulation experimental results demonstrate the promising potential of the adaptive RL for building a human-centered AI system with perceived expressions of empathy to improve dementia care. To further evaluate the system, we plan to conduct real-world user studies.

Keywords. reinforcement learning, intelligent home environment, dementia, human simulator, adaptive intervention, human-centered AI

1. Introduction

Dementia is a progressive neurodegenerative disorder that affects the elderly and diminishes their cognitive functions [1]. As the population ages and caregiver shortages grow, supporting people with early dementia (PwD) to live independently becomes crucial [2]. One of the key challenges is maintaining their circadian rhythm, especially for eating and drinking. Forgetting to eat can lead to serious health concerns. However, it is impos-
sible for caregivers to constantly monitor whether PwD eat enough food. Therefore, an intelligent system integrated with the home environment that can regularly remind and monitor PwD to eat at appropriate times becomes essential. Intelligent home technologies have emerged as promising solutions to enhance the living environment for PwD. However, existing systems often focus on activity detection [3, 4], without considering personalized interventions to provide adaptive supports for eating and drinking scenarios. Meanwhile, previous studies [5, 6] suggest that adaptive interventions are important for PwD, as their needs and body conditions differ from person to person. For instance, visual reminders may not be effective for those with visual impairments. An inappropriate reminder could be frustrating because it shows a lack of empathy for the user's individual needs.

Inzlicht et al. [7] highlight that AI has the potential to offer perceived unconditional empathy without bias, possibly reducing human compassion fatigue, particularly in routine tasks. Therefore, inspired by the just-in-time adaptive intervention (JITAI) [8] design concept, which aims at adapting to an individual’s changing internal and contextual state to provide the right type and amount of support, we introduce the Who takes care artificial intelligence (AI) system. Such a personalized system could adapt to the behavior of the individual user and provide a suitable set of signals to nudge him/her to go to eat. Several unique difficulties were encountered while developing this system. Firstly, as our target users often have cognitive impairments, this can lead to a lack of responsiveness to the system, requiring higher robustness of the system for missing or noisy data. Furthermore, since caregivers indicate that preferences among PwD vary widely due to differing health conditions, the AI must be designed to be perceived as empathic by the user, such that the system can be acceptable and easy-to-use by PwD.

To address these challenges, we design the AI system based on literature [9, 10, 11] and caregivers’ inputs from surveys and interviews. It contains a three-phase framework to provide personalized intervention for nudging PwD to eat and drink during meal times. The reinforcement learning (RL) techniques are employed and integrated into the system to determine a suitable mix of signals for individual users. Before real-world implementation and experiments, we need to ensure the validation of the RL algorithms and the robustness of the AI systems, to avoid errors or unexpected behaviors for our sensitive target users. We therefore further develop a simulated system prototype. Our preliminary study involves simulation experiments to test and refine the system, including how it performs under diverse PwD’s behaviors and how several selected RL algorithms perform under this context. The experiments compare various RL algorithms to identify the most effective one. Additionally, we tested the algorithms’ adaptability and tolerance to the possible changing situation of the environment to evaluate their resilience and flexibility for practical scenarios. This paper aims to develop an AI system that could ensure interventions are not only personalized but also adaptive to individual behaviors and preferences, optimize the Who takes care system in sending right type of signal combinations at the right timing, thus enhancing PwD’s eating and drinking routines in a supportive, human-centric manner.

2. Related Works

Previous studies have investigated the potential of intelligent environment for healthcare to assist PwD in maintaining independent living at home. Several home care technolo-
gies are integrated, including using AI for activity recognition (AR) [12], or utilizing ambient sensors like passive infrared sensors to monitor the residents’ activities. While several studies focus on identifying eating and drinking behaviors [3, 4], others [13, 14] emphasized that relying solely on AR may not fully support PwD. An effective intervention should also remind PwD to go to eat and drink. Moreover, considering the speciality of PwD, the nudging signals should also be designed as personalized and adaptive interventions to fill the gap between the current intelligent home environment and the needs of PwD.

Reinforcement learning (RL) has proven to be an effective and efficient approach for building personalized systems that adapt to individual user behavior and preferences; however, building an application in intelligent environments for PwD introduces new challenges. While several studies [15, 16, 17] have successfully employed RL for personalization in health applications, the scenario could be too complex for users with cognitive impairments who might exhibit a lack of responsiveness. Our project focuses on developing an RL-based system for users with early-stage dementia, particularly to assist in the eating and drinking scenario. By integrating caregiver insights, which we gathered from surveys and interviews, we seek to bridge the gap between technological capability and human-centered design, ensuring that the system is responsive, human-centric, and adaptable to the diverse and evolving needs of PwD.

3. Intelligent Home System

3.1. System Overview

The intelligent home system consists of signal-devices, sensors and an AI module that analyzes the data coming from the sensors and based on them determines which signal-devices to turn on and off. The AI module learns to deliver personalized nudging signals based on user behaviors. We adopted a ‘three-step interaction’ approach [6] into a ‘three-stage escalated eating scenario’ with increasing intensity of signals. As shown in Figure 1, the system reminds PwD three times a day with signals, gently guiding them through three stages. At the end of each stage, sensors detect whether the user has started eating. Both automatic and manual detection methods are employed. For the automatic method, we will use motion or vibration sensors attached to the dining table and chairs to determine if there is an eating activity. For the second method, users will be provided with a controller that allows them to manually indicate ‘I have eaten’ by pressing a ‘Yes’ button. Based on sensor detection and user feedback, if the user has performed an eating activity, the system stops operation; otherwise, after 10 minutes it escalates to a more intensive type of signal to draw the user’s attention. After three iterations, the system stops the current intervention until the next mealtime.

3.2. Reinforcement Learning Method

We formulated our AI module as a contextual multi-armed bandit (CMAB) problem, addressing the unknown preferences of people with dementia in mealtime interventions. Figure 2 represents an overview of the interaction framework, reflecting both the real-world and simulation scenarios. It is worth mentioning that in real-world scenarios,
the simulated UserBot with hidden preferences will be replaced by the human target users. They will interact with the system through ‘Signal-Based Interaction’. The system will present various interventions at different mealtimes. There are six signals, ranging from low to high intensity, including scent (low), music (low), light (medium), image (medium), voice (high), and video (high), based on prior studies on intelligent home technologies [18, 19, 20]. Our interventions comprise eight distinct signal combinations, each consisting of three signals, one from each level of intensity. At each mealtime, the system suggests a three-signal combination, and employs a three-stage Escalation Scenario (as illustrated in Figure 1 for gentle encouragement [6]. In other words, three signals will be delivered one by one with increasing intensity, as described in Section 3.1. We define the key components of RL as follows:

- **Decision Times**: Set $t$ to index decision time: once per meal, three times per day.
- **Contexts**: $C_t \in \mathcal{C}$, time of the day (i.e., breakfast, lunch, and dinner time) is indexed by $C_t$, which indicates the user’s context $C$ at the decision time $t$.
- **Action**: $A_t \in \mathcal{A}$, actions are indexed by $A_t$. There are eight actions (i.e., signal combination) in the action space. In every activation time $t$, the system will choose one action to nudge the user to go eat.
- **Reward**: $R_t \in \{0, 1\}$. After each trial, the system receives a reward from human simulator. $R = 1$ indicates the user reacted or the sensors detected eating activity; $R = 0$ indicates the user did not react or eating activity was not detected.
Six selected algorithms are implemented into the simulation environment: $\varepsilon$-Greedy action selection [21], Upper Confidence Bound (UCB) [22], Thompson Sampling (TS) [23, 24], Deep Epsilon-Greedy [25], Linear Upper Confidence Bound (LinUCB) [26], and Contextual Thompson Sampling (CTS) [24]. Among them, the first three algorithms cannot take the context of action (i.e., time of meal) into consideration for making decisions, while the other three algorithms can. Comparing the performance of various RL algorithms, we aim to not only identify the best-performed algorithm but also investigate the effectiveness of context in our AI system. In this paper, we only demonstrate integrating CTS in the RL framework in Algorithm 1, while the other algorithms follow a similar structure. Beginning with equal initial beliefs, CTS dynamically adapts its approach based on each meal’s context. At decision points, it samples from a Beta distribution, reflecting the probability of each action’s success given past experiences, and selects the most promising action (i.e., signal combination) to display. The CTS beliefs are updated after receiving users’ responses: positive responses strengthen its belief in the action’s success, while negative reactions do the opposite. This continuous learning loop optimizes the nudging strategy and the system eventually displays preferable interventions to the user.

Algorithm 1 Contextual Thompson Sampling Algorithm for Eating Scenario

**Input:** Prior parameters $\alpha, \beta$, contexts $C$, action space $A$

Initialize prior parameters $\alpha, \beta$ to 1 for each context and action

for each decision time $t$ (three times per day for each meal) do

Observe the context $C_t$ (time of day: breakfast, lunch, dinner)

for each action $a$ in $A$ do

Sample $\theta_a(t)$ from a Beta distribution with parameters $\alpha[C_t][a]$ and $\beta[C_t][a]$

end for

Choose action $A_t = \arg\max_a \theta_a(t)$ to nudge the user

Deliver action $A_t$ and observe reward $R_t$

Update $\alpha[C_t][A_t]$ with $\alpha[C_t][A_t] + R_t$ (successes)

Update $\beta[C_t][A_t]$ with $\beta[C_t][A_t] + (1 - R_t)$ (failures)

end for

**Return:** Updated parameters $\alpha, \beta$

### 3.3. Simulation Protocol

Since PwD are particularly sensitive to signals, we introduce a simulation protocol tailored for them, identifying the ethical and practical constraints before deploying the prototype. This protocol employs a human simulator (as depicted in Figure 2), designed to simulate the behaviors and responses of users, allowing us to test our AI system’s effectiveness in a controlled, but also realistic environment. The human simulators are built based on the healthcare literature [9, 10, 11], including the potential responses to the intervention. We further consulted with domain experts about possible response behaviors of PwD, ensuring experiments cover different types of users. Following this protocol, the experiment design cannot only primarily ensure the system’s effectiveness but also align with the needs of PwD before real-world implementation.
4. Simulation Experiments and Results

We evaluate our designed AI system in multiple simulation experiments. To ensure the simulations are aligned with the real situations, we follow the interaction flow presented in Figure 2 and use different human simulator settings to simulate the diverse behaviors of human users. To assess effectiveness, we conducted 600 simulation trials in each simulation run (i.e., three meals per day for six months) with various user preferences for generalizability. Multiple simulation runs were also conducted and the average results over 1000 simulation runs are presented in this section.

4.1. Experiment 0: Contextual and Non-contextual algorithms

We first compared six RL algorithms in the same simulated environment to study the importance of context-based interventions. Each algorithm interacted with a baseline user model having one random specific preferred signal for each meal (e.g. scent for breakfast, light for lunch and image for dinner). As shown in Figure 3, learning curves and average rewards revealed that contextual algorithms CTS, Deep Epsilon-Greedy, and LinUCB outperformed non-contextual counterparts in both effectiveness and speed of learning. CTS achieved the best performance among the algorithms, reaching an average reward close to 1 within 50 trials (i.e., about 17 days’ intervention) and maintaining stability. Furthermore, CTS demonstrated its robustness and reliability through consistently low standard deviations across multiple runs. In the following experiments, we only present the results of three contextual algorithms, which are the best performing algorithms.

![Figure 3. Baseline Results](image)

<table>
<thead>
<tr>
<th>RL Algorithm</th>
<th>Mean (avg. reward)</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epsilon Greedy</td>
<td>0.79 ±0.4</td>
<td></td>
</tr>
<tr>
<td>UCB</td>
<td>0.83 ±0.37</td>
<td></td>
</tr>
<tr>
<td>TS</td>
<td>0.89 ±0.31</td>
<td></td>
</tr>
<tr>
<td>Deep Epsilon-Greedy</td>
<td>0.91 ±0.28</td>
<td></td>
</tr>
<tr>
<td>LinUCB</td>
<td>0.85 ±0.35</td>
<td></td>
</tr>
<tr>
<td>CTS</td>
<td>1 ±0</td>
<td></td>
</tr>
</tbody>
</table>

4.2. Experiment 1: User Type Adaption

We investigated how well different RL algorithms could adapt to diverse user preferences in a simulated environment designed to manage mealtimes for PwD. According to domain experts, we created five user types (A-E) with varying responses to nudging signals: Type A positively reacts to up to two signals per meal (e.g. music and light for breakfast, light and video for lunch, and music and scent for dinner); Type B is the same as Type A but will never react to one meal (e.g. no reaction during lunch); Type C positively responds only to one signal type (e.g. video for all three meals); Type D shows no reaction to any signal; and Type E positively reacts to one of the signals in a certain intensity level (e.g. music for breakfast, and scent for lunch and dinner).
Figure 4 showed that all RL algorithms could learn and adapt to user preferences, achieving stable average rewards across user types A, C, and E. CTS algorithm generally outperformed the others in these cases. Interestingly, for user type C with unchanging preferences, LinUCB exhibited a faster learning rate than CTS, suggesting its potential for scenarios where the relationship between context and reward is unclear.

4.3. Experiment 2: Preference Shifts Over Time

This experiment evaluates the AI system’s ability to dynamically adapt to users’ shifting preferences, which is important for maintaining engagement with users whose needs and behaviors may change over time. If so, we aim to further identify which RL algorithm can learn and quickly converge in this dynamic scenario. Thus, we introduce two user types: Type F, where users initially respond to one type of nudging signal but then suddenly shift to another after one month (i.e., 90 trials), and Type G, depicting users who become unresponsive over time. In User Type F (results in Figures 5a), if the user maintains responsiveness to the signal, the RL algorithm can adapt to the shift within about 50 trials (meals). For User Type G (results in 5b), the loss of responsiveness may provide a warning sign to caregivers to intervene and indicate potential changes in the user’s dementia status.

4.4. Experiment 3: Acceptance & Tolerance Test

We aim to further evaluate how the RL algorithms can perform under realistic and challenging conditions. The Acceptance Test examined the algorithm performance when user responses do not consistently translate into rewards, introducing acceptance rates rang-
(a) User Type C, Acceptance Rate = 0.75
(b) User Type C, Acceptance Rate = 0.5
(c) User Type E, Tolerance Rate = 0.2

Figure 6. Experiment 3: Acceptance & Tolerance Test

ing from good (0.75) to low (0.25) to mimic the unpredictable nature of real-world scenarios. Figure 6a, 6b indicated a significantly declined performance as acceptance rates decreased, highlighting the system is sensitive to the environment’s response.

Tolerance Test evaluated system robustness against noisy environments by incorporating 10% and 20% invalid data to simulate errors in user input or sensor data. Figure 6c showed moderate resilience to data inaccuracies, with certain impacts on performance as error rates increased. Together, these experiments demonstrate the robustness of our AI system and emphasize the importance of maintaining user engagement and reducing noisy data from the home environment.

5. Conclusion & Future Work

In conclusion, this paper presents a novel, RL-based AI system, specifically focusing on eating and drinking scenarios for PwD in an intelligent home environment. Our simulation experiments tested various RL algorithms to determine their effectiveness in a simulated environment mimicking real-world situations, including users’ behaviors and noisiness. The results showcase the successful simulation of a human-centered AI system, demonstrating the potential of the RL system in adapting to the unique behaviors and needs of PwD, in order to support them to live independently. For future work, we aim to extend our research by conducting real-world trials with early-stage PwD and further validate our system more thoroughly. Although we have cooperated with domain experts and caregivers to make the AI human-centered, based on the current small range of scenarios, the empathicness of the AI is still limited. With deeper investigation into the interaction between humans and AI, we aim to enhance our system’s capabilities. Our eventual goal is to create a personalized, adaptive, and empathic AI system that can meet the needs of PwD and release the burden of their caregivers.

Acknowledgement

We would like to extend our special thanks to Regieorgaan SIA for funding this research (RAAK.PUB09.060) and all the consortium partners of the Wie Zorgt project.
References


Hybrid Intelligence for Digital Humanities

Victor DE BOER 1 and Lise STORK
Vrije Universiteit Amsterdam, the Netherlands.

Abstract. In this paper, we explore the synergies between Digital Humanities (DH) as a discipline and Hybrid Intelligence (HI) as a research paradigm. In DH research, the use of digital methods and specifically that of Artificial Intelligence is subject to a set of requirements and constraints. We argue that these are well-supported by the capabilities and goals of HI. Our contribution includes the identification of five such DH requirements: Successful AI systems need to be able to 1) collaborate with the (human) scholar; 2) support data criticism; 3) support tool criticism; 4) be aware of and cater to various perspectives and 5) support distant and close reading. We take the CARE principles of Hybrid Intelligence (collaborative, adaptive, responsible and explainable) as theoretical framework and map these to the DH requirements. In this mapping, we include example research projects. We finally address how insights from DH can be applied to HI and discuss open challenges for the combination of the two disciplines.

Keywords. Hybrid Intelligence, Digital Humanities, Cultural Heritage

1. Introduction

Digital tools and methods have permeated all fields of science, including the humanities. This has given rise to the area of scholarly activity known as Digital Humanities (DH)[26]. With more museum, archive and library collections made available in digital form, either through digitization or as ‘digital born’, questions arise on how digital tools of exploration, analysis, annotation and visualisation can be incorporated in the humanities workflows and methodologies. Here, it has been recognized that the use of such tools can make this research more effective and efficient and open up new possibilities for analysing data or telling stories. At the same time, these tools and interdisciplinary methods will need to be carefully matched to the values of the humanities research process with respect to transparency and reproducible methods [4].

More recently, Artificial Intelligence (AI) tools and methods have become more available also to stakeholders in the field of humanities. Researchers from both AI and humanities have been identifying potential challenges and opportunities for AI in Digital Humanities. These opportunities include the possibility to analyse heritage and other humanities data at scale [22], open up new ways of modelling humanities knowledge, analysing data and recognizing patterns, but also disseminating results [41]. Challenges identified in literature include issues around bias in data and/or AI algorithms [5,36].

1Corresponding author: Victor de Boer, v.de.boer@vu.nl
as well as challenges and ethics of categorization [29], or the lack of polyvocality in representation [19]. Concerns have also been expressed around the issue of diversity, equity and inclusion (DEI) in the larger field of AI itself [12]. Such investigations call for more human-centric approaches to AI to be developed that match Digital Humanities requirements. We argue here that the recently introduced research paradigm of Hybrid Intelligence [1] aligns well with these requirements.

In this paper, we explore the synergies between DH as a discipline and HI as a research paradigm. We identify that in Digital Humanities research, the use of digital methods and specifically that of AI is subject to a set of requirements and constraints that are well-supported by the capabilities and goals of Hybrid Intelligence. We take the CARE principles of Hybrid Intelligence (collaborative, adaptive, responsible and explainable) as theoretical framework and map these to the DH needs of tool and data criticism, transparency and ethical considerations. We also address how insights from DH can be applied to challenges identified in the context of Hybrid Intelligence, specifically where it concerns developing responsible and transparent systems. Through a description of example DH projects, we further illustrate this synergy. We finally identify open challenges and opportunities for the combination of the two disciplines.

2. Challenges for the use of AI in Digital Humanities

2.1. Understanding Digital Humanities

Digital humanities (DH) represents a scholarly realm where computing or digital technologies intersect with the various disciplines within the humanities. This field encompasses the methodical utilization of digital resources in humanities research, coupled with the examination of their practical applications[17]. DH is characterized by innovative approaches to scholarly endeavors, involving collaborative, transdisciplinary, and computationally oriented practices in research, teaching, and publishing[10]. By introducing digital tools and methodologies into humanities studies, DH acknowledges the shift away from the printed word as the primary medium for knowledge production and dissemination. DH researchers employ a diverse set of methodologies and tools to analyze, interpret, and present data. Textual analysis stands as a cornerstone, leveraging computational techniques such as natural language processing (NLP) and sentiment analysis to dissect literary works, historical documents, and cultural texts. Additionally, DH embraces network analysis, utilizing graph theory to map relationships among entities, uncovering hidden connections within datasets, and shedding light on complex societal structures. Visualization tools play a pivotal role, translating intricate data into visual representations that aid in comprehending patterns, and relationships within information.

2.2. AI in DH

Within Digital Humanities, there has been quite a lot of interest in the application of Artificial Intelligence technologies for annotation, analysis and dissemination of DH research. Various types of AI solutions are being applied to specific DH challenges. Knowledge Representation approaches have been used to capture and model data, information and knowledge from various domains. Especially Semantic Web principles and practices.
have been embraced by the galleries, archive, libraries and museum (GLAM) community, to connect heterogeneous and distributed collections[27,14]. By using standard data models, interoperability is increased and standard ways of reasoning can be applied. At the same time, representation of heritage data as knowledge graphs also allows for other downstream analyses such as through machine learning or visualisation. Machine Learning has been used to analyse large humanities datasets, with a recent focus on exploring deep learning approaches. Challenges for the use of AI in DH also include the limited extent to which current Machine Learning solutions can be used for multimodal or sparse datasets (e.g. [33,44]). Critical views on the role of Machine Learning in DH have been part of scholarly debate, focusing on bias, transparency and the way that usage of ML could influence the methodological rigour of humanities research [3]. The humanities focuses on artifacts produced by human agents, and the majority of these artifacts have historically been natural language texts. Natural Language Processing (NLP) therefore takes an important role in DH. Most NLP pipelines now contain at least some Machine Learning, either using pre-trained (large) language models (LLMs) or using models specified for the data and task at hand [46]. For other modalities, specific approaches exist. For example, for visual artifacts Computer Vision is used for analysis, information extraction, as well as reconstruction and restoration. While traditionally, more feature extraction-based pipelines were used, with the recent success of (deep) neural networks for this task, such approaches are now applied to humanities datasets, leading to novel insights, but also exposing challenges around its usage [48].

2.3. Requirements and challenges

As argued above, since its early beginnings, DH has not only concerned itself with the application of digital methods to enhance humanities research, but also has focused on a critical reflection of data, tools, and workflows. Especially in the usage of AI tools and methods, the way that such tools collaborate with (human) scholars has been questioned. Collecting and categorizing these elements allows us to define five challenges for successful integration of AI tools in DH. We list them below in short form and expand on the challenges and their mapping to HI capabilities in Section 4.

Embedding in scholarly practice. AI should be embedded in the DH research methodology, through collaboration between human and artificial agents.

Data or source criticism. Humanities data have specific characteristics that makes them different from other domain data. Moreover, a critical stance to data is a key component in the humanities and successful AI needs to support this.

Tool criticism. Similarly, a critical attitude towards methods and tools used is necessary. A successful AI systems supports critical inquiry into itself.

Polyvocality and diversity. The humanities recognizes that multiple perspectives are present and the diversity of (citizen) scientists, data and voices is key.

Distant vs Close reading. In the digital humanities there is ongoing debate about the relation between distant reading (shallow analysis at scale) versus close reading (deep inquiry of text). This presents excellent opportunities for hybrid solutions.
3. Hybrid Intelligence and the CARE principles

Hybrid Intelligence as a field concerns itself with the investigation and design of human-AI ecosystems. As such, it builds on AI research, but focuses on interaction with users. [1] recognizes four research challenges, abbreviated as the CARE principles. These principles can be formulated as capabilities for AI systems that are effective in HI scenarios:

**Collaborative.** AIs systems will need to be designed and constructed in such a way that they are able to cooperate in synergy with human actors in a variety of tasks. This collaboration should ultimately take into account the strengths and limitations of both the human and artificial agents.

**Adaptive** HI systems function in dynamic environments with diverse human-agent teams, necessitating adaptability to changing contexts, variable team structures, preferences, and roles.

**Responsible** Mitigating risks requires integrating ethical and legal considerations into the design and operation of HI systems. Values like transparency, accountability, privacy, and fairness must be intrinsic to the design process and performance.

**Explainable** For mutual understanding between intelligent agents and humans, explanations are vital for shared awareness, goals, and collaborative strategies. This requires the use of methods and algorithms that are not only transparent, but can interact with users to explain their reasoning.

4. Mapping CARE principles to DH requirements

In this section, we take the requirements and challenges as identified in Section 2 and explore how the CARE principles introduced in Section 3 can be applied to these challenges. Each time we present a description of the challenge; which of the HI principles are most suited to be mapped to this case and open challenges.

4.1. Embedding in the Scholarly Practice

In order to build good AI tools for DH that users can and will effectively adopt in their daily practice, creators of such tools need to understand the strengths and weaknesses of humans and AI alike. This requires them to adapt to scholars’ preferences in working with and gaining understanding of historical material, and their goals, cultural biases, and social norms. Common roles that scholars take in such collaborative efforts are that of the reader or analyst, that of the annotator in crowd or niche-sourcing [16] endeavors, or that of the curator of enriched material. In their workflow, humanities scholars tend to prefer close over distant reading of source material (see Section 4.5), and tend to avoid systematic keyword searches. Moreover, they often work exploratory; in search of interesting pieces of information that contribute to the point they aim to make, and might need to consult sources of hundreds of years old [47]. AI systems can tailor to such needs, but creators of such tools also require understanding of how humanities scholars perceive the output of AI systems. Errors, for instance, can create skepticism towards the capabilities of tools. For example, whereas error rates of 60-95% for a handwriting recognition system can greatly improve searchability of archival content, Schomaker et
al. mention that “Scholars [...] are disappointed by the OCR quality of machine-printed text, which may be far above 95” [43]. Moreover, humanities researchers tend to be more skeptical towards sophisticated tools, and prefer easy of use and transparency [23].

In an effort for natural dialogue between AI and humanities scholars, DH has started using eXplainable AI (XAI) techniques to generate insights or explanations of humanities datasets [18,15]. Examples may take the form of narratives, i.e., [6,13] in which data are organized according to important historical events and actors, mimicking how humans understand phenomena. Moreover, interpretable rules can be used to explain associations. Wilcke et al. [49], for instance, mine explainable associations from archaeological materials as novel insights for archaeologists. Such XAI tools remain limited in the humanities [18], but have the potential to increase interpretability of collections, as well as generate trust in human-machine collaboration.

**Open challenges** Which cognitive biases do scholars have in understanding humanities datasets and AI systems? What are natural ways of interacting via AI systems?

### 4.2. Data or Source Criticism

Data or source criticism deals with the critical evaluation of sources in the context of historical studies. This includes assessing the quality, completeness, bias, and representativeness of the data, as well as considering the ethical implications of data collection and use. These concern not only texts, but also other modalities such as maps, photographs and drawings [30]. For humanities scholars it is obvious that historical inequalities lead to datasets that are incomplete, biased, hide certain perspectives (see Section 4.4) such as indigenous people’s voices or certain social groups [40,19,2], or use language that is inappropriate in today’s society [9,37].

With use of ML, especially generative AI models, data criticism becomes even more important. Humanities datasets have typically not been created as ML benchmarks, but for the purpose of knowledge communication, preservation, correspondence, or other. Moreover, they are highly heterogeneous, and their structure changes over time [2]. How can AI systems act responsibly, when built on top of untrustworthy sources? Other than building responsible datasets, AI systems should be capable of explaining the sources on which they base their reasoning in a transparent way.

Knowledge representation can help organize data, providing insights on knowledge gaps. Datasheets [2], for instance, describe datasets in a structured way, so that these can be fit more easily to the information needs of data re-users. Moreover, adhering to principles of FAIR data management [51] will not only increase the reusability of digital data and methods, but also ensure that they can be subjected to the same rigorous criticism of tools and data as is common in humanities research [31]. However, when applying standard data models and reasoning for the interpretation of historical sources, there is the risk of simplification or reinterpretation of such sources in the current scientific paradigm, as socio-cultural data is inherently vague and ambiguous [29].

**Open challenges** What is the best way to measure whether AI systems act responsibly with respect to the data that they use? Moreover, how can AI systems explain to humanities scholars the provenance of data used for explanations or insights?
4.3. Tool Criticism

Related to data criticism is tool criticism. Defined in [20] as "[...] the critical inquiry of knowledge technologies considered or used for various purposes [...]". As such, tool criticism is at the heart of the DH research methodology. It calls for critical reflection on how tools came to be, how the tool functions and the way in which digital tools are used in the digital history methodology[31]. As an example, when using NLP tools, an uncritical with respect to the tools capabilities recognizing spelling variations or dealing with diacritical characters could lead to incorrect conclusions. Specifically for AI technology, the danger exists that these tools are more opaque: black-box models have a hard time explaining why certain results are produced, risking perpetuating existing biases [40] and large language models do not always allow for tracing query results back to original sources. In order to support a critical view on the use of tools in a hybrid setting, AI systems need to be transparent about their inner workings, what data they are trained on, what their limitations are and how small changes in settings influence results. This requires that notions of data provenance are integrated in such tools[32,38]. The capabilities of responsible maps well onto this requirement as systems that collaborate with humans need to be developed in responsible ways (responsible in design). They also need to be aware of and behave in a responsible manner, specifically in being transparent. This also makes it clear why explainable systems are needed: When systems are able to communicate about their inner workings, they allow for the required critical inquiry by the humanities scholar. A special mention must be given to Generative AI, that produces novel output based on patterns learned from data. Use of this relatively new form of AI is starting to be explored in the DH context. Examples include using Generative Adversarial Networks to analyse visual styles [39] or generate images of medieval manuscripts depicting steam trains[11]. Retrieval Augmented Generation (RAG) combines an information retrieval component with text generation based on large language models[21]. When responses or 'alternative' texts, images or other sources can be generated convincingly, it becomes even more important that such methods are transparent.

Open challenges. One of the main challenges here is how state-of-the-art ML can be made 'criticisable'. Here the field of explainable AI (XAI) can provide opportunities, but for successful integration, more research is needed into how DH scholars can interact with systems in order to retrieve suitable explanations. At the same time, standards for describing and sharing tool descriptions are needed, building on [31].

4.4. Polyvocality and Diversity

In the humanities, the concept of positionality is often used to indicate that the research process itself as well as the researcher are not neutral, but bring their own perspective and biases[8]. Polyvocality is essential in digital humanities as it promotes a diverse and inclusive approach to the study of cultural artifacts and human experiences in the digital realm[42,19]. By incorporating a variety of voices and perspectives, polyvocality helps mitigate biases, avoid stereotypes, and reflect the complexity of the real world. It enhances interdisciplinary collaboration, engages communities, and contributes to more comprehensive representations of cultural heritage[24]. Where platforms facilitate public engagement, polyvocality becomes crucial for fostering a more inclusive dialogue about
culture, history, and society. Overall, embracing multiple voices in digital humanities enriches research, analysis, and understanding, leading to a more nuanced exploration of the complexities inherent in human narratives and cultural phenomena. This can refer not only to diversity in the collection, but also of the persons involved in the research activities, including experts and lay persons. This need for polyvocality connects well with the goals of responsible AI, specifically where it concerns fairness towards various representations and the way in which systems deal with bias. In DH practice, this has for example the implication that methods used for annotating datasets (be they automatic, through crowdsourcing or hybrid solutions) will need to be able to capture, represent and present such diverse perspectives[19]. This requires that user-facing methods are aware of and can deal with diverse end-users. This diversity can be in the expertise dimension, but also in the cultural or age dimension. Collaborative AI is needed, specifically AI that is able to collaborate with a variety of users. In that sense it will have to be adaptive to be able to deal with multiple perspectives, various cultural contexts and be able to keep ‘disagreement’ between experts, laypersons and algorithms intact.

Open challenges Diversity of users will be a key challenge for successful HI tools that support or even promote polyvocality. Many AI systems and interfaces are designed for specific end-users, be they domain experts or laypersons. They often are accessible only to a limited number of people, and might require considerable resources. Systems that cater to diverse users will be needed to make sure all relevant voices can be heard.

4.5. Distant and Close Reading

Taken from literary studies, Distant reading and close reading represent two contrasting approaches to humanities research. Close reading involves meticulous examination and interpretation of individual texts, focusing on details like language, imagery, and structure to uncover deeper meanings and themes. Conversely, distant reading employs methods to analyze large corpora of texts, aiming to identify broader patterns, trends, and relationships across works. It emphasizes more quantitative analysis, data visualization, and the exploration of phenomena at scale[35]. While traditional humanities research often includes much close reading, with the rise of computational methods, in DH, more interest for distant reading has emerged for a variety of modalities[28]. Recognizing this complementarity paves the way to use collaborative AI to the fullest effect: HI systems will need to collaborate with human scholars and be flexible enough to switch between supporting distant reading, but also supporting close reading. They will need to be adaptive in the sense that the results of close reading will be needed to be taken into account for further processing. Pattern recognition using Machine Learning can significantly enhance this distant reading and connecting large numbers of sources through knowledge graphs allows for querying and analysis of large interconnected corpora. Human experts can then engage more deeply with smaller amounts of objects of study. For example, the CLARIAH media suite specifically is designed to support both this distant reading (using AI to connect audiovisual content) and close reading (providing easy access and annotation tools for individual videos)[34]. Where now, these activities are disparate, [7] recognize that further exploitation of the complementarity between computational methods and humans is a worthwhile research direction. Such deeper collaboration between humans and machines fits the research agenda and capabilities of HI well.
HI mapping. Collaborative, Adaptive

Open challenges. To support the seamless transition between distant and close reading, AI systems need to maintain connections between aggregations and the original sources. LLMs for example, often only store the statistical aggregations and cannot trace back query answers to specific texts. Knowledge Graphs that connect various data sources now often only include metadata, and not the actual documents, images or videos. Integral solutions are needed that actually contain such multimodal data (e.g. [50]).

5. Discussion

In the preceding section, we have outlined how the capabilities of HI match with the requirements of Digital Humanities. HI can make deep collaborations between AI systems and with human scholars possible. It can be truly successful if such systems are designed taking into account the diversity of the users and methodological concerns that they have concerning tools and data. This calls for systems that deal well with provenance and transparency and are responsible by design.

While we argue here that DH as a field can benefit from the Hybrid Intelligence initiative, we also recognize that the humanities provide HI not just as a playing ground for interesting and difficult challenges regarding human-AI collaboration, but also that for fundamental AI/HI challenges, the field can learn much from the humanities. Questions such as how to deal with multiple perspectives, how to conceive of and deal with social bias and (algorithmic) fairness should not be answered solely by AI experts, but can be addressed more deeply by taking into account the decades of experience that exists in the humanities addressing such issues. With the increased impact of AI in society and the volume of current public debate about its ethics, AI in general, and HI in specific, is in need of humanists “are ideal “domain experts” for the current juncture” [25].

As Hybrid Intelligence is a current research initiative, many of the mappings identified above are future research directions. However, we here end our paper with two specific overarching challenges towards HI for DH. First of all, a danger in Digital Humanities is that of one-shot solutions. Tools and methods are often designed for a single (research) project, and lack in reusability across collections or subfields. There should be a focus on reusable, Open Source, tools, datasets and methods that are well described (see Tool and Data criticism) and usable for a variety of end-users. This usability also requires careful development of user-interfaces that are well-integrated in the humanities task at hand. This can include web-portals, dialogue systems or interactive notebooks [45].

Secondly, although we here have been mostly focusing on how HI systems should be designed, of course the human scholars will need to play their part. To have effective Hybrid teams, the human actors will also need to be able to be critical and cooperative with their computational team mates. This requires both an open mind, and increased AI literacy. Including digital methods in the humanities curricula will be crucial to ensure that the next generation of humanities scholars is ready to play their part in the HI team.

Acknowledgements. This research was supported by the Hybrid Intelligence Center, funded through the Netherlands Organisation for Scientific Research (NWO), https://hybrid-intelligence-centre.nl (024.004.022) and by the Pressing Matter project (http://pressingmatter.nl), funded by NWO through the Netherlands National Research Agenda (NWA.1292.19.419).
References


Gibbs, F., Owens, T.: Building better digital humanities tools: Toward broader audiences and user-centered designs. Digital Humanities Quarterly 6(2) (2012)


Haigh, T.: We have never been digital. Communications of the ACM 57(9), 24–28 (2014)


Kitamoto, A., Nishimura, Y.: Data criticism: a methodology for the quantitative evaluation of non-textual historical sources with case studies on silk road maps and photographs. In: Third Annual Conference of the Japanese Association for Digital Humanities (JADH2013). pp. 11–12 (9 2013)


Schomaker, L.: Design considerations for a large-scale image-based text search engine in historical manuscript collections. it-Information Technology 58(2), 80–88 (2016)


EASY-AI: sEmantic And compoSable glYphs for representing AI systems

Alexis ELLIS a, Brandon DAVE a, Hugh SALEHI a, Subhashini GANAPATHY a and Cogan SHIMIZU a
a Wright State University

Abstract. Despite the rapid integration of artificial intelligence (AI) into various research domains and the lives of everyday people, challenges with communicating and understanding these AI systems arise. The lack of a consistent method of communication highlights the need for a transdisciplinary approach to explain the inner workings of AI systems in a cohesive and accessible manner. We thus propose an ontological visual framework using semantically-enhanced symbols, providing a symbolic language for conveying the structure, purpose, and characteristics of AI systems. The framework encompasses a generalizable glyph set of various AI system components, ensuring both common and obscure architectures can be represented. In this paper, we present the underlying logical formalisms that dictate the behavior of this visual framework as a means to significantly enhance the comprehensibility and understandability of AI system behaviors.

Keywords. Human-Computer Interaction, Semantics, Ontology, Symbology

1. Introduction

As artificial intelligence (AI) systems surge into not only the spotlight of major research interest in different domains but also its rapid integration into our daily lives, so too does the need to understand, explain, and interpret exactly what an AI system is doing. As the complexities of these systems increase with the state of the art, the finer details of their functionality may become locked behind opaque systems. It would behoove this area of research to create a standardized means of communicating the structure of AI systems comprehensively to users of various backgrounds. One approach is to develop a visual framework that utilizes symbols to convey the complexities of AI systems. Providing a visual framework of AI systems using symbols has many benefits since symbols have been used by humans since our earliest recorded history. Presently, the use of symbols, more notably known as emojis, are used in ways that convey entire sentences, feelings, or thoughts. Moreover, symbols have the potential to convey information cross-culturally. Therefore, we have built upon the state of the art and created a visual framework that aims to communicate and represent AI systems, this framework is comprised of sEmantic And compoSable glYphs (EASY-AI). EASY-AI conveys a visual of the various constraints that govern components of an AI system in an intuitive and easily comprehensible way for the user. EASY-AI will provide a means for more effective communication and better comprehension of AI systems with varying complexities. EASY-AI aims to ease not only the tensions between AI and its integration into society.
but also the lack of understanding that surrounds it by providing a clear and concise way of communication thus increasing human-computer interaction for the future.

2. The EASY-AI framework

This section introduces the breakdown of our tool EASY-AI from transformational ontology design pattern(s) taken from the Boxology format introduced by [15] to an underlying formalized ontology. We demonstrate EASY-AI through 3 distinct layers: 1 The visual framework layer corresponds directly to the structures indicated in Boxology. 2 The classification layer explains the 1-to-1 transformation from the top visual layer into concepts for ontological manipulation. 3 EASY-AI’s framework will be presented in the final layer. We represent this 3-layer perspective to demonstrate how EASY-AI was constructed and how this tool works at a fundamental level. We end our description with a use case to demonstrate EASY-AI in a real-world application.

2.1. The Boxology

The Boxology patterns from [15] provide the visual foundation to simplistically break down the inner workings of an AI system, but lack any formalization to dictate how the Boxology patterns interact with one another. These patterns are considered our first layer, providing a user-friendly, surface-level glimpse of what an algorithm might be doing internally. However, the idea is to provide a visual guide to AI Systems with an in-depth look at the processes that take place within these patterns. The core takeaway from Boxology is that each symbol represents a component of an AI system. The application of EASY-AI will take a specific interest in the core building patterns, also referred to as the elementary patterns. Since EASY-AI is an extension of Boxology we will be adopting the taxonomy to keep the information consistent. A brief description of the fundamental taxonomy is as follows: Instances: are represented by rectangles. Processes: are represented by ovals. Models: are represented by hexagons. Actors: are represented by triangles. Solid arrows: indicating the sequential order of the flow of the input. The taxonomy outlined above represents the extent of our discussion on Boxology research. However, should you be interested, the paper is accessible in our reference section. EASY-AI takes another more in-depth look at the intricacies that lie under these patterns without losing the intuition the Boxology visually provides to the user.

2.1.1. Ontological Transformation

The EASY-AI framework implements an ontology which is the representation of concepts and interlinked relationships with specifications defining how these connections are allowed to functionally operate together. This second layer is included as an auxiliary layer for demonstrating how the transition from Boxology to ontology is being conducted to eliminate the potential of another "black box phenomenon." The ontological transformation is a 1-to-1 translation of the patterns into conceptualized classes, where the dashed red lines visually represent this 1-to-1 translation. In our example, using patterns 1a and 1b depicted in Figure 1, though both have different defined inputs, at the EASY-AI layer these features are generally the same, reiterating the importance of keeping the Boxology patterns as a visual component for a quick and comprehensible first assessment of an AI system before diving deeper into its intricacies.
2.2. EASY-AI

The final layer from Figure 1, represented in the previous section, is where we reveal our in-depth observations of the underlying components that the Boxology visual patterns are conveying. This section will delve into the specific concepts and roles that make up the rules that govern EASY-AI.

2.2.1. Formalization

The EASY-AI framework is seeded with a core pattern from modular ontology [10], specifically the Data Transformation pattern [13]. The full formalization of the EASY-AI framework is graphically depicted in Figure 3a. In this schema diagram, we represent conceptual classes as gold rectangles, while blue dashed rectangles represent a class described externally from EASY-AI [13]. Filled, directed arrows represent binary relations between classes and open arrows represent subsumption. For visual convenience, the graphical representation includes grey boxes to group similar entities; however, the grey boxes have no ontological meaning.

As the name suggests, the DataTransformation pattern is an established representation of the transformation process from Boxology to ontology. We implement Process as an abstraction of DataTransformation concepts. From the pattern, the notion of Roles, which describes the part that a particular concept plays in some transformation activity, can be established and assigned to concepts that are accepted and produced by a Process. Thus, OutputRole represents the role for concepts produced by Process. However, Boxology has functional usages within multi-step tasks that require the product of a Process to frequent the notion of playing the part of InputRole for the next Process. EASY-AI incorporates the notion of Roles to be assigned to Boxology instances.

2.2.2. Axioms

The EASY-AI framework provides both a schema to describe concept connections via relationships and also implements axioms, or constraints, to describe the shape of how concepts and relationships adhere. EASY-AI incorporates 8 axioms, as depicted in Figure 2, which is a condensed list as provided by [11]. The axioms constrain the shapes to which a relationship \( R \) is applied between the concepts \( A \) and \( B \). **Axiom 1** represents a subsumption relationship to describe that properties of Concept \( A \) are inherited from the parenting Concept \( B \). **Axiom 2** asserts scoped domain of \( R \), scoped by \( A \), is \( B \). **Axiom 3** asserts scoped range of \( R \), scoped by \( A \), is \( B \). **Axiom 4** asserts that for every \( A \), there has to be a relationship \( R \) with \( B \). **Axiom 5** asserts that for every \( B \), there has to be an inverse...
relationship of $R$ with $A$. Axiom 6 asserts that for every $A$, there is at most one relationship $R$ with $B$. Axiom 7 asserts that for every $B$, there is at most one inverse relationship of $R$ with $A$. Axiom 8 asserts that an $A$ may have a relationship $R$ with $B$. This axiom is all encompassing for the whole ontology, as it is explicit to represent the possibility of existence.

These are the broad constraints that preside over the data transformation axioms that we have adopted. Boxology instances with their relationships are described within the EASY-AI framework in Sections 2.3-2.5. Boxology processes with their relationships are described within Sections 2.6-2.13.

2.3. Actors

The concept of Actor can be defined as an entity, whether human or otherwise (i.e., software), which engages in actions or behaviors within a system or environment. Actor-performsInputRole-InputRole: An Actor can only ever perform as an InputRole to a Process. This is constrained by Axioms 3, 4, 6, and 7. Actor-participatesInProcess-Engineer: An Actor is always present with the Process abstracted to Engineer. This is constrained by Axioms 4, 5, and 7.

2.4. Artifacts

EASY-AI unifies the concept of Data and Symbol to be implemented with the notion of Artifacts, as the abstracted concepts have shared relationships between other concepts. As with [2], the distinction of Data and Symbols, as defined in [1], are still followed for EASY-AI. Data can be represented similar to other AI systems, which include numbers, text, tensors, and streams (representing a sequence of data). Symbols are distinct from Data as Symbols represent labels, relationships between concepts, and records reflected by Data.

Artifact-participatesInProcess-(Transform or Train or Deduction): Artifacts can participate in either Processes of Transform, Train, and Deduction. The relationship of participatesInProcess with concepts of Model and Train or Deduction are not constrained by any axioms.

Artifact-performsInputRole-InputRole: An Artifact can perform as an InputRole to a Process. This is constrained by Axioms 3, 4, 6, and 7.

Artifact-performsOutputRole-OutputRole: An Artifact can perform as an OutputRole to a Process. This is constrained by Axioms 3, 4, 6, and 7.

(Data and Symbol) SubClassOf-Artifact: Data and Symbol inherit from Artifact. This is represented by Axiom 1.
2.5. Models

In the same nature as Artifacts, a top-level description is used to describe Models. As described in [2], models are descriptions of entities and their relationships.

Model-participatesInProcess-(Train or Deduction): A Model is a participant of either Processes of Train or Deduction. The relationship of participatesInProcess with concepts of Model and Train or Deduction are not constrained by any axioms.

Model-performsInputRole-InputRole: An Model can only ever perform as an InputRole to a Process. This can be constrained with Axioms 3, 4, 6, and 7.

Model-performsOutputRole-OutputRole: An Model can only ever perform as an OutputRole to a Process. This can be constrained with Axioms 3, 4, 6, and 7.

SemanticModel-SubClassOf-Model: Models can be abstracted for distinction; such as, the representation of SemanticModels, implemented with Axiom 1. SemanticModels represent the meaning of symbols identified by their concepts, and relationships [14].

2.6. DataTransformation

EASY-AI inherits the DataTransformation pattern, as previously stated in Section 2.2.1. The axioms that govern these are:

DataTransformation-providesRole-(InputRole or OutputRole): providesRole is adopted by the DataTransformation pattern. This relationship can connect to either an InputRole or OutputRole concept, which will be later assigned to a respective Boxology instance.

As described in [13], the Roles allow for the distinction of sourced data to retain its representation after an applied process or transformation, as these tend to be destructive to the originating sourced data. The axioms applicable to this relationship are: Axioms 2, 4, 5, 6, and 7.

DataTransformation-implements-Process: A DataTransformation can be implemented with a particular Process. The relationship of implements between the concepts of DataTransformation and Process can be applied with Axioms 2, 3, 5, and 7.

2.7. Process

As defined in [15], Processes represent a series of actions or operations to an end.

Process-SubClassOf-DataTransformation: Process inherits from DataTransformation. This is represented by Axiom 1.

2.8. Generate

Generate acts as an abstraction of a Process. The axioms that govern these are:

Generate-SubClassOf-Process: Generate inherits from Process. This is represented by Axiom 1.

2.9. Engineer

Engineer only occurs when a Process involves an Actor. The axioms that govern these are:

Engineer-SubClassOf-Generate: Engineer inherits from Generate. This is represented by Axiom 1.

Engineer-accepts-Actor: The Process of Engineer can only accept Actors. This is constrained by Axiom 2, 3, 4, 5, 6, and 7.
2.10. Train

Generate acts as an abstraction of a Process. The axioms that govern these are:
Train-SubClassOf-Generate: Train inherits from Generate. This is represented by Axiom 1.
Train-accepts-Artifact: Processes of Train can only accept Artifacts. This is constrained by Axioms 3, 4, and 6.
Train-produces-Model: Processes of Train can only produce Models. This is constrained by Axioms 2, 4, and 6.

2.11. Infer

Infer acts as an abstraction of a Process. This process typically results in a conclusion or opinion based on known facts [15]. The axioms that govern these are:
Infer-SubClassOf-Process: Infer inherits from Process. This is represented by Axiom 1.
Induction-SubClassOf-Infer: Induction inherits from Infer. This is represented by Axiom 1.
Deduction-SubClassOf-Infer: Deduction inherits from Infer. This is represented by Axiom 1.
Deduction-accepts-(Artifact and Model): Processes of Deduction only occurs when the InputRoles are a joint combination of Artifact and any Model.

This is constrained by Axioms 2, 3, 4, 5, 6, and 7.
Deduction-produces-Symbol or Model: Processes of Deduction can either produce Symbols or Models. The relationship of produces with concepts of Deduction and Symbol or Model are not constrained by any axioms.

2.12. Transform

Transform acts as an abstraction of a Process. The axioms that govern these are:
Transform-SubClassOf-Process: Transform inherits from Process. This is represented by Axiom 1.
Transform-produces-Data: Processes of Transform can only produce Data. This is constrained by Axioms 2, 3, 4, 5, 6, and 7.

2.13. Embed

Embed is a particular kind of Transform process if, and only if, the Transform process accepts Symbols and SemanticModels and produces Data.
Embed-SubClassOf-Transform: Embed inherits from Transform. This is represented by Axiom 1.
Embed-(accepts-Symbol and SemanticModel/produces-Data): Processes of Embed only occurs when the InputRoles are a joint combination of Symbol and SemanticModel and the OutputRole is performed by Data. This is constrained by Axioms 2, 3, 4, 5, 6, and 7.
Embed-produces-Data: Processes of Embed only provides an OutputRole to Data. This is constrained by Axioms 2, 3, 4, 5, 6, and 7.
2.14. Availability and Licenses

The ontology documentation, which includes the serialization of the ontology into Turtle syntax and each schema diagram for every pattern in [15], is provided online via a GitHub repository. These resources are licensed permissibly under the MIT license.

3. Use-Case Scenario

To represent EASY-AI’s capabilities and real-world application, we have applied our framework to a Roomba use-case explored in [3]. This Roomba utilizes a Reinforcement Learning (RL) core with 2 attached sensors allowing navigation throughout an office space. The EASY-AI framework for this use-case can be seen in Figure 4. To simplify the construction of this Roomba in EASY-AI’s framework, by not reiterating the 1-to-1 translation, we have presented our rendition of this use-case with color-coded labels: Blue boxes represent the ontological representation of Boxology instances as Input, as described in Section 2.1.1. Purple boxes correspond to the ontology representation for Output. Yellow boxes correspond to the ontology translation of Boxology processes as Process. These colors act as a guide as this is the first introduction to EASY-AI and its application from theory to the real world. The Roomba use-case puzzles together the patterns to represent the base use-case; however, the modularity of EASY-AI allows us to piece the reinforcement learning aspect into the robotic devices’ process providing a clear and detailed formalization of the internal workings of the Roomba under the Boxology patterns. This use-case serves as a working example of how we can pair the Boxology patterns as an acting top layer of observation for a user, fulfilling the visual aspect of representing an AI system, while having a second layer of EASY-AI that ontologically displays the intricacies and relationships more descriptively but staying easily tied to the visuals of the Boxology patterns in a real-world application.

https://github.com/kastle-lab/easy-ai

(a) Graphical representation of EASY-AI ontology. (b) Graphical representation of transforming Boxology elementary pattern 2c to an ontology design pattern.
4. Conclusion

The EASY-AI framework provides an underlying formalization for the graphical structures from the already established visual Boxology. The goal of EASY-AI is to provide an in-depth inspection of the concepts and rules that govern the individual aspects we see in the Boxology patterns. This in-depth inspection will allow for a more established comprehension of AI systems and more effective communication between not only expert users in collaboration but also everyday users. EASY-AI will provide the grounds for a healthy synergy in human-machine interaction.

Future Work There are several foreseeable next steps with this framework. We anticipate the connection of additional ontological resources, including the provenance ontology [9] and computational observation and environment patterns [12,4] for significantly improved descriptions of AI software systems. furthermore, we now must also connect the actual glyphs to underlying formalisms herein described. Finally, we will leverage EASY-AI building tools similar to Orange [7], where the logical constraints can accelerate auto-complete-like functionality when constructing, analyzing, and visualizing AI systems.

References


Effective Maintenance of Computational Theory of Mind for Human-AI Collaboration

Emre ERDOGAN\textsuperscript{a} and Frank DIGNUM\textsuperscript{b} and Rineke VERBRUGGE\textsuperscript{c}
\textsuperscript{a} Utrecht University, Utrecht, Netherlands
\textsuperscript{b} Umeå University, Umeå, Sweden
\textsuperscript{c} University of Groningen, Groningen, Netherlands

Abstract. In order to enhance collaboration between humans and artificially intelligent agents, it is crucial to equip the computational agents with capabilities commonly used by humans. One of these capabilities is called Theory of Mind (ToM) reasoning, which is the human ability to reason about the mental contents of others, such as their beliefs, desires, and goals. For an agent to efficiently benefit from having a functioning computational ToM of its human partner in a collaboration, it needs to be practical in computationally tracking their mental attitudes and it needs to create approximate ToM models that can be effectively maintained. In this paper, we propose a computational ToM mechanism based on abstracting beliefs and knowledge into higher-level human concepts, referred to as \textit{abstractions}. These abstractions, similar to those guiding human interactions (e.g., trust), form the basis of our modular agent architecture. We address an important challenge related to maintaining abstractions effectively, namely abstraction consistency. We propose different approaches to study this challenge in the context of a scenario inspired by a medical domain and provide an experimental evaluation over agent simulations.

Keywords. Theory of Mind, Abstraction, Human-AI Collaboration, Human-inspired Computational Model

1. Introduction

Hybrid Intelligence (HI)\cite{1} is about integrating human and machine intelligence for the purpose of expanding human intellect instead of replacing it. Effective HI requires human-agent collaboration at its core, where a human and a computational agent partner, working together, complement each other’s abilities to create fruitful partnerships.

One of the capabilities that help humans successfully maintain their social interactions with other humans is called Theory of Mind (ToM) reasoning\cite{2,3}. Put simply, ToM refers to the human capacity to reason about the mental content of others such as their beliefs, desires, values, and goals. Possessing a functioning ToM is critical to understand and predict others’ behaviour\cite{4} and can provide further benefits to the possessor when used recursively. Recently, many computational ToM models have been developed to understand its effectiveness in competitive, cooperative, and mixed motive settings\cite{5,6,7,8,9,10,11,12}. The results are mostly positive, suggesting that utilizing ToM leads to enhanced performance in the studied tasks.
Developing a practical computational ToM for human-agent interaction is valuable but challenging. Many existing ToM-using agent models begin by representing individual beliefs about others and constructing a ToM model from these. In contextually rich settings that feature continuous interaction, the agent accumulates numerous beliefs about others over time, some of which are applicable only in specific contexts and others are useful in different situations. This makes developing a comprehensive ToM model infeasible since computationally tracking all individual mental attitudes of all others is a costly approach. To continue being effective in its interactions with human partners over time, the agent should be efficient in keeping, maintaining, and utilizing these beliefs.

One candidate solution to this problem comes from human behaviour, called abstracting. As a problem solving technique, abstracting enables humans to form a broad understanding of the problem and its potential solutions, rather than focusing on specific details [13]. In complex social settings, this approach helps us approximate what we should look for in the interaction to reach our goals and make our decisions accordingly. Consider trust as an abstraction, which serves as a backbone in collaboration and mainly captures agents’ confidence in each other’s abilities, reliability, and commitment [14]. A human being, by using the abstracting technique, can efficiently utilize the relevant information about their partner to decide whether to trust the partner or not. Coupled with ToM reasoning, the human being can further understand if the partner trusts them back and correspondingly decide which actions they need to perform in their interactions.

This paper proposes a computational ToM mechanism based on abstracting beliefs and knowledge into higher-level abstractions that serve as practical approximations. We design a formal agent architecture based on epistemic logic [15] that provides a modular structure for storage and maintenance of individual beliefs, knowledge, and abstractions. For this agent to work with humans collaboratively, we need to address a challenge regarding abstraction consistency: Since an agent’s beliefs and knowledge change over time, it is necessary to devise methods to revise abstractions from time to time in an efficient manner. We propose different mechanisms to study this challenge in the context of a scenario inspired from a medical domain and provide evaluation over simulations.

The rest of this paper is organized as follows. Section 2 sets up our motivating human-agent collaboration example. Section 3 describes how we formalize abstractions and computational ToM reasoning with epistemic logic for computational agents. Section 4 illustrates the use of the formal agent design, addresses our solutions to the challenge given above, and evaluates our solutions over agent simulations. Section 5 discusses our work in relation to related work and provides pointers for future directions.

2. Motivating Example

We consider a motivating example from the medical domain, featuring a collaboration between a computational agent doctor X and a human doctor Y for diagnosing a patient Z (inspired by the work of Erdogan et al. [16]). In this scenario, the computational agent doctor X is designed to complement the capabilities of the human doctor Y to improve the efficiency of the diagnostic process. They can distribute the tasks based on their respective strengths throughout the diagnostic procedure [17]. For instance, Y can handle tasks that require human traits such as conducting patient interviews and performing physical examinations, while X can focus on tasks that are computationally more viable [18,19] such as diagnostic testing (e.g., medical imaging [20]). In this setup, which can be seen
as a collective decision-making process, both doctors can share their results with each other and discuss the diagnosis together.

We are particularly interested in situations where there is a conflict between decisions of $X$ and $Y$ and social skills are being used to resolve such conflicts through different techniques [21,22]. We particularly focus on trust. By aggregating beliefs and knowledge that are contextually relevant, $X$ can determine whether it should trust $Y$ or not. What makes it more interesting is that $X$ can also reason about how $Y$ abstracts her knowledge and beliefs to decide whether to trust $X$ or not (i.e., how $Y$ does her own approximation for trust) with the help of its computational ToM of $Y$. These abstractions can help $X$ in choosing the best response to go with when a dispute occurs.

In this setting, $X$ needs to explicitly have the contextually relevant beliefs and knowledge that are necessary to create and maintain its trust-related abstractions. These knowledge and beliefs can be generated and revised internally (e.g., by different processes of the agent) or acquired from external sources. $X$ also needs a mechanism to do the required approximations properly. In our scenario, we assume that for $X$ to trust $Y$, it needs to know that $Y$ is a doctor, and believe that $Y$ communicates well with $X$, and believe that $Y$ is an expert in her field. Moreover, as part of ToM reasoning, $X$ should also be able to capture $Y$’s understanding of trust as that can be different than itself. We assume that $X$’s ToM model for $Y$ depicts that for $Y$ to trust $X$, she needs to believe that $X$ has good diagnostic capabilities and good communication skills.

3. Formal Design

3.1. Abstraction Elements

We define agent as an entity that has beliefs and knowledge about other agents, maintains its beliefs and knowledge over time, and uses them when interacting with other agents. Our formalization is based on epistemic logic [15] where propositions can be created from propositional atoms, together with negation and conjunction operators and knowledge and belief modalities per agent.

**Knowledge and Beliefs:** To formally represent knowledge and beliefs of a set of agents $\mathcal{X}$, we use the following language $\mathcal{L}_{KB}^{\mathcal{X}}$ given by the Backus-Naur form:

$$\varphi ::= p \mid \neg \varphi \mid \varphi \land \varphi \mid K_{X} \varphi \mid L_{X} \varphi$$

Here, $p$ are propositional atoms and $X \in \mathcal{X}$. Given $p_1 =$ “$Y$ is a doctor”, $K_{X} p_1$ and $L_{X} p_1$ can be read as “the agent $X$ knows that $Y$ is a doctor” and “the agent $X$ believes that $Y$ is a doctor”, respectively. Notice that $L_{Y} K_{X} p_1$, which states that “the agent $Y$ believes that the agent $X$ knows that $Y$ is a doctor”, is a member of $\mathcal{L}_{KB}^{\mathcal{X}}$. Formulas with nested epistemic operators are useful in expressing agents’ higher-order knowledge and beliefs.

**Abstractions:** An abstract concept $A$ is a human-inspired, abstract decision-making heuristic which can guide agents in their interactions. $\mathcal{A}$ denotes the set of abstract concepts in our framework. Essentially, these abstract concepts are meaningful when defined in a relational manner. Thus, we define an abstraction as a proposition structured as $A(X,Y)$, where $A \in \mathcal{A}$ and $X,Y \in \mathcal{X}$.

Our formalization allows agents to hold (higher-order) knowledge and beliefs about abstractions in a nested manner. For example, given $p_2 = \text{Trust}(X,Y)$, $K_{Y} p_2$ and $L_{X} L_{Y} p_2$
can be read as “the agent Y knows that X trusts Y” and “the agent X believes that the agent Y believes that X trusts Y”, respectively. We refer to such (higher-order) knowledge and beliefs (about abstractions) simply as abstractions.

Abstractions can come about in different ways. For example, there is a vast literature on how one agent can learn if it should trust another agent using machine learning techniques [23,24]. The use of machine learning techniques necessitates to have a large number of interactions before a decision can be obtained. However, in real life, usually one needs to decide to trust another agent without the opportunity to interact with her too many times. Social cues and organizational constructs help in determining a quick decision. For example, you might trust someone because they are a doctor in a reputable hospital, even though you had no previous interactions with her. At the same time, it is important to be able to identify the reasons that led to trust [25], which is difficult to explain with machine learning techniques. Since our focus is on enabling agents to create abstractions quickly and reason on them effectively, we formalize abstraction through predefined rules, rather than data-driven techniques.

**Abstraction Rules**: An abstraction rule is a derivation rule in the form of \( \varphi \rightarrow \phi \) such that \( \phi \) is an abstraction. For instance, \( p_1 \rightarrow p_2 \) (i.e., “Y is a doctor” implies that “X trusts Y”) and \( K_X p_1 \rightarrow L_X K_Y p_2 \) (i.e., “X knows that Y is a doctor” implies that “X believes that Y knows that X trusts Y”) are both regarded as abstraction rules. \( \varphi \) could refer to various roles as described above as well as external information (e.g., my friend trusts Y).

Epistemic logic is useful for formally exploiting the implications of various epistemic principles such as \( K_X p \rightarrow p \) (i.e., what is known is true) and \( K_X p \rightarrow K_X K_X p \) (i.e., what is known is known to be known). We use the following prominent epistemological principles \( P_J \), \( P_K \) and \( P_L \) in tandem with abstraction rules to derive abstractions:

\[
\begin{align*}
P_J &: K_X \varphi \rightarrow L_X \varphi \quad \text{(i.e., knowledge implies belief)} \\
P_K &: K_X (\varphi \rightarrow \phi) \rightarrow (K_X \varphi \rightarrow K_X \phi) \quad \text{(i.e., knowledge is closed under implication)} \\
P_L &: L_X (\varphi \rightarrow \phi) \rightarrow (L_X \varphi \rightarrow L_X \phi) \quad \text{(i.e., belief is closed under implication)}
\end{align*}
\]

For example, say we have the knowledge \( K_X (p_1 \rightarrow p_2) \) (i.e., X knows that “Y is a doctor implies that X trusts Y”) and \( K_X p_1 \) (i.e., X knows that Y is a doctor). By using \( P_K \) and modus ponens, we can derive that \( K_X p_1 \rightarrow K_X p_2 \) and hence, \( K_X p_2 \).

3.2. Agent Architecture

Our proposed agent architecture consists of three modules. The **knowledge and belief module** keeps the agent’s knowledge and beliefs. At certain times, the agent derives abstractions from these, which are stored in the **abstraction module**. The **deliberation module** uses the abstractions to make decisions on how to interact with other agents.

**Knowledge and Belief Module**: For an agent X, \( X^{KL} = (M,N) \) represents X’s **knowledge and belief module** such that:

- \( M \) is the knowledge set of X such that every member of the set is either of the form \( K_X p \) (first-order knowledge), \( K_X K_Y p \), or \( K_X L_Y p \) (second-order knowledge) where \( p \) is a proposition and \( Y \in \mathcal{X} \), and
- \( N \) is the belief set of X such that every member of the set is either of the form \( L_X p \) (first-order belief), \( L_X K_Y p \), or \( L_X L_Y p \) (second-order belief) where \( p \) is a proposition and \( Y \in \mathcal{X} \).
Whenever the agent interacts with others, the information that reaches the agent is stored in this module. $M$ is a dynamic set such that new knowledge can be added to it. As generally understood in the literature, we assume that knowledge is always true. Hence, new knowledge would not conflict with existing knowledge, making $M$ conflict-free by definition. On the other hand, $N$ contains beliefs, which may or may not be true. Thus, a new belief can easily conflict with an existing one. For simplicity, we ensure that a newly created belief overrides the older conflicting belief, making the set conflict-free.

**Abstraction Module:** For an agent $X$, $X^{Abs} = (O, P, Q)$ represents $X$’s abstraction module such that:

- $O$ is $X$’s abstract concepts where $O \subseteq \mathcal{O}$,
- $P$ is $X$’s abstraction rules such that every member of the set is of the form $K_X(\varphi \rightarrow \phi)$ or $L_X(\varphi \rightarrow \phi)$ where $\varphi \rightarrow \phi$ is an abstraction rule, and
- $Q$ is $X$’s (current) abstractions.

We assume that $O$ is a static set that contains all possible abstract concepts the agent can have over time. Note that this set might be large but that does not mean that the agent will have abstraction rules in $P$ or existing instances of abstractions in $Q$ related to them. Each agent has a set of abstraction rules in $P$, which can be different for each agent. While we do not necessarily focus on the rules themselves, it is possible that this set is dynamic such that new rules are added as the agent sees fit or some rules are removed if not seen fit. $Q$ holds the abstractions which the agent derives by using $P_J$, $P_K$, $P_L$, and modus ponens on its beliefs and knowledge. Similar to $N$, it is conflict-free: newer values of abstractions override the older values.

**Deliberation Module:** For an agent $X$, $X^{Del} = (R, S, T)$ represents $X$’s deliberation module such that:

- $R$ is the set of actions that $X$ can do when interacting with other agents,
- $S$ is action deliberation rules of $X$ such that every member of the set is of the form $\phi \rightarrow r$ where $\phi$ is an abstraction or a conjunction of abstractions and $r \in R$, and
- $T$ is $X$’s (current) action where $T \in R$.

We assume $R$ is a static set, such that possible actions cannot change during execution. $S$ defines how the agent will deliberate with others based on its abstractions. For this paper, we assume these are given for each agent but essentially these rules can evolve over time based on interactions with others. For simplicity, we assume deliberation rules clearly identify which action will be taken in a given situation. $T$ keeps the current action.

We represent $X$ as $X = (X^{KL}, X^{Abs}, X^{Del}) = ((M, N), (O, P, Q), (R, S, T))$. Although separated conceptually, the modules are connected to each other functionally. For example, for the agent to maintain its abstractions in $P$, it needs to check the content of $X^{KL}$. Also, for the agent to decide on its next action, it uses the action deliberation rules in $S$ in combination with $P$. The next section illustrates this flow within our motivating example.

### 4. Abstraction Consistency

To illustrate use of abstractions in human-agent collaboration, we formalize the conflict resolution scenario given in Section 2 in which $X$ needs to decide on the action to take next depending on the set of abstractions that $X$ has about $Y$. For the
sake of simplicity, we limit the number of possible abstractions that \( X \) can use to two, namely \( L_X(\text{Trust}(X,Y)) \) and \( L_X(\text{Trust}(Y,X)) \) (i.e., \( X \)'s trust towards \( Y \) and \( Y \)'s trust towards \( X \), respectively) and the number of possible actions to four, namely \textbf{Converse}(\( Y \)), \textbf{Agree}(\( Y \)), \textbf{Persuade}(\( Y \)), and \textbf{Consult}(\( Y \)). Table 1 describes the content of all modules of \( X = \langle (M,N), (O,P,Q), (R,S,T) \rangle \) and shows how \( X \) can use them.

**Table 1.** Design in use: \( X \) decides on its next action according to its abstractions.

<table>
<thead>
<tr>
<th>( M )</th>
<th>( k_X(\text{Doctor}(Y)) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>( L_X(\text{Expert}(Y)) )</td>
</tr>
<tr>
<td></td>
<td>( L_X(\text{GoodCommunication}(Y)) )</td>
</tr>
<tr>
<td></td>
<td>( L_XL_Y(\text{GoodCommunication}(X)) )</td>
</tr>
<tr>
<td></td>
<td>( L_XL_Y(\text{GoodCapabilities}(X)) )</td>
</tr>
</tbody>
</table>

\( O \) \{Trust\}

\( P \) \( L_X(\text{Doctor}(Y) \land \text{Expert}(Y) \land \text{GoodCommunication}(Y) \rightarrow \text{Trust}(X,Y)) \)

\( L_XL_Y(\text{GoodCommunication}(X)) \land L_Y(\text{GoodCapabilities}(X)) \rightarrow \text{Trust}(X,Y) \)

<table>
<thead>
<tr>
<th>( Q )</th>
<th>( L_X(\text{Trust}(X,Y)) )</th>
</tr>
</thead>
</table>

\( R \) \{Converse, Agree, Persuade, Consult\}

<table>
<thead>
<tr>
<th>( S )</th>
<th>( L_X(\text{Trust}(X,Y)) \land L_X(\text{Trust}(Y,X)) )</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>( \text{Action} )</th>
<th>( \text{Converse}(Y) )</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>( \text{Action} )</th>
<th>( \text{Agree}(Y) )</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>( \text{Action} )</th>
<th>( \text{Persuade}(Y) )</th>
</tr>
</thead>
</table>

| \( \text{Action} \) | \( \text{Consult}(Y) \) |

\( T \) \textbf{Converse}(\( Y \))

The goal of \( X \) is to interact with \( Y \) according to the abstractions that it has about \( Y \). To do that, it first derives the abstractions in \( Q \) by using \( M \), \( N \), \( P \), and epistemological principles \( P_T \), \( P_K \), and \( P_\land \). Then, by using \( Q \) and \( S \), it decides on the action to take next. Table 1 shows that both \( L_X(\text{Trust}(X,Y)) \) and \( L_X(\text{Trust}(Y,X)) \) hold at the beginning; so, \( X \) chooses to resolve the conflict by conversing with \( Y \) to arrive at a joint resolution.

\( X \) may need to resolve many other conflicts during its partnership with \( Y \). Thus, \( X \) needs to choose the correct actions through its abstractions that are consistent with its current knowledge and beliefs. To do that, \( X \) needs to update \( Q \) in accordance with the changes in \( M \) and \( N \). Furthermore, the update mechanism should be as efficient as possible for \( X \) to be effective in its action-decision process: \( X \) should update \( Q \) as much as needed, but also, only when necessary. Since \( M \) and \( N \) can also change with knowledge and beliefs that do not affect the derivation of abstractions, \( X \) should be careful not to do any unnecessary update-checks for its abstractions. Thus, \( X \) requires a concrete procedure for deciding when to revise its abstractions and the accompanying actions.

### 4.1. Update Methods

In order to realize abstraction consistency, an agent needs to check and update its abstractions. There are a number of factors that are important to consider, such as frequency (of updates), change (in the knowledge and beliefs), and engagement (with others). Considering these factors, we formulate the following update strategies (in \textbf{bold}):
• **Frequent** updates its abstractions after every change in $N$, without considering engagement.
• **Infrequent** updates its abstractions after every 10 rounds, without considering if there are changes or engagement.
• **Revision** updates its abstractions after every belief revision in $N$, but does not consider belief addition or engagement.
• **Deliberation** updates its abstractions only before deliberation without considering changes or frequency.
• **Change** updates its abstractions only before deliberation and only if there is a change in $N$, without considering frequency.
• **Selective-Change** updates its abstractions only before deliberation and only if there is an abstraction-related change in $N$.

**Hypothesis 1** An agent $X = \langle X^{KL}, X^{Abs}, X^{Del} \rangle$ with **Selective-Change** Strategy is the most effective (among all six strategies) in obtaining abstraction consistency.

### 4.2. Evaluation

We evaluate the performance of these strategies over simulations. We have designed a computational agent to simulate the behaviour of $X = (\langle M, N \rangle, \langle O, P, Q \rangle, \langle R, S, T \rangle)$ for the human-agent collaboration scenario. $X$ is capable of adding beliefs to and revising beliefs in $N$, updating its abstractions in $Q$, and doing deliberations. We do not simulate human doctor $Y$’s behaviour explicitly in the simulations.

A simulation lasts 10000 rounds. There are three types of events that can occur with the same probability (1/3) during the simulation: i) an ordinary (i.e., non-abstraction-related) belief is created and $X$ adds it to $N$ (or revises it in $N$); ii) an abstraction-related (e.g., $L_X(\text{Expert}(Y))$) belief is created and $X$ adds it to $N$ (or revises it in $N$); or iii) a deliberation moment comes (i.e., conflict occurs) and $X$ decides on the action to take.

The experiment provides two basic metrics to measure agent performance. One is the number of abstraction updates an agent does throughout the simulation. The second one is the number of consistent abstractions that it has at the time of deliberations, where an abstraction is considered consistent at the time of deliberation if current knowledge and belief base of the agent would also infer the same abstraction. For example, suppose $X$ chooses Converse($Y$) because of having $L_X(\text{Trust}(X, Y))$ and $L_X(\text{Trust}(Y, X))$ in $Q$ but in fact $L_X(\text{Expert}(Y))$ is recently removed from $N$. This implies $L_X(\text{Trust}(X, Y))$ should not be in $Q$ as well, implying a mistake; hence, makes it an inconsistent abstraction. We measure abstraction-effectiveness by combining these two metrics, which is equal to the number of consistent abstractions used in deliberations per abstraction update.

We have run the simulation 10 times with 10 different seeds for randomization purposes and averaged the results. Table 2 shows the number of times abstractions are updated as well as the number of times correct abstractions are used in deliberations (for each agent type). As expected, **Frequent** agent performs the most updates as it does so with every change in its belief set and thus has zero errors (e.g., its abstractions are always up-to-date). **Infrequent** agent, on the other hand, performs the fewest updates but because of this it had the most errors in its abstractions. **Revision** agent is less efficient than Infrequent agent, yet it also has a lower usage of inconsistent abstractions. **Deliberation** agent and **Change** agent do not create inconsistent abstractions but they are also not as economical as Infrequent agent in their abstraction updates. **Selective-Change**,
Table 2. Abstraction-effectiveness of 6 strategies over 10 simulations. Average number of deliberations: 3305

<table>
<thead>
<tr>
<th>Agent Type</th>
<th>#Abstraction Updates</th>
<th>#Consistent Abstractions</th>
<th>Abstraction-Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent</td>
<td>3401.1</td>
<td>3305</td>
<td>0.97</td>
</tr>
<tr>
<td>Infrequent</td>
<td>1000</td>
<td>2584.7</td>
<td>2.58</td>
</tr>
<tr>
<td>Revision</td>
<td>3296.1</td>
<td>3302.1</td>
<td>1.00</td>
</tr>
<tr>
<td>Deliberation</td>
<td>3305</td>
<td>3305</td>
<td>1.00</td>
</tr>
<tr>
<td>Change</td>
<td>1668.2</td>
<td>3305</td>
<td>1.98</td>
</tr>
<tr>
<td>Selective-Change</td>
<td>1110.1</td>
<td>3305</td>
<td>2.98</td>
</tr>
</tbody>
</table>

on the other hand, excels in the task. It does not make mistakes when using abstractions in deliberation moments and is nearly as effective as Infrequent agent in updating its abstractions. Among all, Selective-Change agent illustrates the best use of abstractions in deliberations, being the most abstraction-effective; this corroborates Hypothesis 1.

5. Discussion

There exist recent agent models that use various forms of computational ToM reasoning for human-agent collaboration. Piazza and Behzadan [26] design a ToM-based technique to differentiate agents based on cooperativeness where communication plays a crucial role. Wu, Sequiera, and Pynadath [27] focus on understanding how humans interact in collaborative teaming settings where they know little about others’ goals and intentions. Bara et al. [28] introduce the concept of collaborative plan acquisition, where humans and AI agents work together to learn and communicate to acquire a complete plan for joint tasks. Their results highlight the importance of modeling a partner’s mental states explicitly. Montes et al. [12] introduce an agent model that combines ToM reasoning with abductive reasoning capabilities and demonstrate their computational ToM model’s performance in the context of the Hanabi game [29]. Erdogan et al. [16] provide an outline of a computational ToM framework based on abstracting individual beliefs into higher-level concepts such as social roles, norms, and human values.

Our work provides a novel approach in ToM-based agent modeling with explicit use and maintenance of abstractions, yet it is also constrained by several limitations. Firstly, an abstraction-using agent should model that different people can build and maintain their abstractions in different manners (e.g., their trust [30]), which our examples do not feature. Thus, to be more flexible in collaborations, such an agent should be capable of properly updating its abstractions when they are not working well (e.g., changing its abstraction rules concerning others’ abstractions such as their trust in itself). Moreover, we only include the concept of trust in our examples and evaluation so that we sufficiently focus on the mechanics of the abstraction use and maintenance. Including other abstractions, such as respect or affinity, can enable us to create more detailed models.

Acknowledgements

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research (https://www.hybrid-intelligence-centre.nl).
References


Towards Effective Management of Verbal Probability Expressions Using a Co-Learning Approach

Christian FLEINER\textsuperscript{a,b,c,1} and Joost VENNEKENS\textsuperscript{a,b,c}

\textsuperscript{a}KU Leuven, De Nayer Campus, Dept. of Computer Science
\textsuperscript{b}Leuven.AI – KU Leuven Institute for AI, Belgium
\textsuperscript{c}FlandersMake@KULeuven – M&A, Belgium

ORCiD ID: Christian Fleiner https://orcid.org/0000-0003-4280-670X, Joost Vennekens https://orcid.org/0000-0002-0791-0176

Abstract. Domain experts are one of the most important knowledge sources when building a knowledge base. However, communication about uncertain states and events is prone to misinterpretations and misunderstandings, because people prefer to convey probability estimations by verbal probability expressions (VPEs) which have a high between-subject variability. Additionally, several biases exist when expressing uncertainty verbally. Nevertheless, the application of VPEs might be necessary. Therefore, means must be identified to manage VPEs and to translate them into numeric values appropriately. In this paper, we propose a co-learning approach with example to efficiently and effectively communicate (subjective) probabilities of states and events in teams where human and AI team members are familiarized with the translation between VPEs and numeric values until both parties are capable of using solely numeric values.

Keywords. human-agent collaboration, hybrid intelligence, hybrid team, knowledge acquisition, preference paradox, subjective probability, uncertainty communication

1. Motivation

Domain experts are one of the most important knowledge sources when building a knowledge base (KB). However, knowledge acquisition remains one of the major challenges in many domains (e.g., in the field of knowledge representation and reasoning (KRR) [1]). The quality of elicited expert knowledge is heavily influenced by the means of communication. In reference to the Shannon-Weaver communication model [2], misinterpretations and misunderstandings between sender and receiver can be caused by inefficient encoding and decoding, but also by communication interferences (noise). Knowledge about uncertain states or events is especially challenging to communicate due to ambiguities, approximations that incorporate vagueness, coarseness, or simplifications, and the fact that estimates of likelihood of states and events depend on the acquired sample and natural randomness [3, Table 1.12]. After all, an expert’s shared

\textsuperscript{1}Corresponding Author: Christian Fleiner, christian.fleiner@kuleuven.be.
opinion is nothing more than “a subjective assessment, evaluation, impression, or estimation of the quality or quantity of something of interest that seems true, valid, or probable to the expert’s own mind” [3, p. 98].

Many experts prefer to use verbal probability expressions (VPEs) to convey estimations of likelihood. The application of VPEs is addressed by cross-disciplinary research in the field of psychology, cognitive science, intelligence analysis, climate analysis, but also industrial engineering. (Non-exhaustive) synonymous terms are verbal probabilities, probabilistic phrases, judgment terms, verbal uncertainty expressions and words of estimative probabilities. We adopt the term verbal probability expression which was used by Beyth-Marom [4] and adopted by Teigen [5]. In this paper, we first review the history of research on VPEs and then apply the learned insights to uncertainty communication in a human-agent team. One widely agreed on hypothesis is the so-called preference paradox. The preference paradox describes that a sender prefers to use words to communicate uncertainties while a receiver prefers to receive numeric estimations about uncertainties [6,7]. While the preference paradox might be cumbersome between human experts, a collaborating human-agent team could naturally react to the preferences as the software agent (or robot) “thinks” numerically and can be enhanced with natural speaking capabilities.

Additionally, the human-agent team cannot only react, but also diminish the preference paradox by engaging into co-learning [8,9]. In the co-learning scenario, software agents learn the set of VPEs that individual human team colleagues use and use this information to assign (objective) probability ranges to the expressions. Furthermore, software agents can regularly provide feedback to human team members to improve their understanding of their individual verbal-to-numeric probability mapping until all team members feel comfortable using numeric values. In this matter, we make the following three contributions:

- Summarize the main results of research done in the field of VPEs
- Provide an example to show the potential of using VPEs in human-agent teams
- Outline future research directions including two research questions

2. Literature overview

We roughly distinguish research on VPEs in two trend waves. A central element of the second wave (2013 – now) is the empirical research on numerically bounded linguistic probability (NBLP) schemes [10]. As the second wave is still ongoing, an overview of NBLP schemes is provided in this section and relevant insights of the wave are discussed in section 4.


In the beginning of research on VPEs, several studies (predominantly in the linguistic and medical field) were conducted to infer the numeric interpretation of expressions [4,11,12,13,14,15,16] (non-exhaustive). The set of expressions varies among studies containing single adjectives, adverbs or nouns (e.g., rare, likely, and toss-up). Additionally, word combinations with modifiers (e.g., very likely), hedges (e.g., almost never),
Figure 1. Illustration of the structures of established NBLP schemes. All probability ranges are mapped with their label in Table 1. The ranges were optimized for visibility and should thus not be used as direct reference. For instance, “impossible (A)” of the WEP scheme refers to 0%, but is displayed as the range between 0–1%. Also, overlapping ranges like IPCC’s and EFSA’s “more likely than not” (> 50%) are removed.

Table 1. Overview of verbal probability expressions (VPEs) used by numerically bounded linguistic probability (NBLP) schemes with reference (R) to Figure 1 indicating the probability ranges. The VPEs are ordered by ordinal rank from lowest (A) to highest (I).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Impossible</td>
<td>Exceptionally unlikely</td>
<td>Almost no chance</td>
<td>Remote</td>
<td>Highly unlikely</td>
<td>Almost impossible</td>
<td>Remote chance</td>
</tr>
<tr>
<td>B</td>
<td>Almost certainly not</td>
<td>Extremely unlikely</td>
<td>Very unlikely</td>
<td>Highly improbable</td>
<td>Unlikely</td>
<td>Extremely unlikely</td>
<td>Highly unlikely</td>
</tr>
<tr>
<td>C</td>
<td>Probably not</td>
<td>Very unlikely</td>
<td>Unlikely</td>
<td>Improbable (improbably)</td>
<td>Even chance</td>
<td>Very unlikely</td>
<td>Unlikely</td>
</tr>
<tr>
<td>D</td>
<td>Chances about even</td>
<td>Unlikely</td>
<td>Roughly even chance</td>
<td>Roughly even odds</td>
<td>Likely</td>
<td>Unlikely</td>
<td>Realistic possibility</td>
</tr>
<tr>
<td>E</td>
<td>Probable</td>
<td>About as likely as not</td>
<td>Likely</td>
<td>Probable (probably)</td>
<td>Highly unlikely</td>
<td>About as likely as not</td>
<td>Likely (probable)</td>
</tr>
<tr>
<td>F</td>
<td>Almost certain</td>
<td>Likely</td>
<td>Very likely</td>
<td>Highly probable</td>
<td>-</td>
<td>Likely</td>
<td>Highly likely</td>
</tr>
<tr>
<td>G</td>
<td>Certain</td>
<td>Very likely</td>
<td>Almost certain (ly)</td>
<td>Nearly certain</td>
<td>-</td>
<td>Very likely</td>
<td>Almost certain</td>
</tr>
<tr>
<td>H</td>
<td>-</td>
<td>Extremely likely</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Extremely likely</td>
<td>-</td>
</tr>
<tr>
<td>I</td>
<td>-</td>
<td>Virtually certain</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Almost certain</td>
<td>-</td>
</tr>
</tbody>
</table>
negations (e.g., not unreasonable), and phrases (e.g., liable to happen) were considered. The number of VPEs varied between 12 [13] and 52 [15] expressions.

In summary, the results of the studies indicate that subjects can consistently rank VPEs on an ordinal scale. The maximal number of discriminable probability ranks appears to be seven in accordance with results from Miller [17]. The between-subject variability of probability estimations in all studies, however, was conspicuous indicating the nonexistence of an unconditioned consensus. [18]

Another research track focused on epistemic modalities. The terms possible, probable and certain [19] can be ranked on an ordinal scale without assigned numeric probability ranges. Also, speakers’ expression of uncertainty about the truthfulness of a statement is composed of the speakers’ belief that the statement is true and their confidence in their belief [20,21,22]. Therefore, in the context of the probabilistic estimation of uncertain events, we also must distinguish between the estimated probability and the speakers’ confidence.

2.2. Overview of NBLP schemes

NBLP schemes are pre-defined templates that translate a set of VPEs to numeric values. Multiple organizations have developed their own NBLP schemes (based on results of the first research wave) to make qualitative expert judgments easier to compare and reduce the potential of misinterpretation. In the following, we outline six established NBLP schemes. The structures of the NBLP schemes are illustrated in Figure 1. The expression sets of the schemes are displayed in Table 1.

- The Intergovernmental Panel on Climate Change (IPCC) introduced a first version of its calibrated language in 2005 [23] which was refined in 2007 [24, p. 23]. The refined version contains the expressions “extremely (un)likely”. Its application was more described in detail in 2010 [25]. The expressions are referred to as likelihood terms. The purpose of the calibrated language is to facilitate communication among the working groups of the IPCC. Experts are encouraged to complement any likelihood assessment with an confidence level on an ordinal 5-point scale that ranges between very low confidence and very high confidence. The scheme has the option to say “more than 50%” which overlaps with other categories. The expressions are defined as thresholds and the scheme itself should not be used when listing facts [25].

- The NBLP scheme from the European Food Safety Authority (EFSA) Scientific Committee is derived from the IPCC scheme [26, p. 61]. The main difference is that the expression “more likely than not” was separated from the other expressions to prevent overlapping ranges. Also the IPCC expression “exceptionally unlikely” was renamed to “almost impossible”, assuming to avoid misinterpretations to the similar sounding expression “extremely unlikely”. While the IPCC scheme interprets VPEs as thresholds, the EFSA scheme interprets them as probability ranges. Another important feature is that the EFSA scheme explicitly permits experts the option to abstain from assigning a probability due to missing information or other reasons.

- The Intelligence Community Directive (ICD) 203 [27, p. 3] introduced another NBLP scheme with the goals to enable objective judgments, mitigate bias, and provide assessments in time. The scheme offers two alternative expression sets.
One set mainly modifies the expression “likely”, the other set modifies the expression “probable”. The ICD scheme illustrates the interchangeability of the words “likelihood” and “probability”. Additionally, it addresses the issue that the effect of modifiers is yet unclear by using for one set the modifier “very” and for the other set the adverb “highly”. Items of the expression sets should also not be mixed to avoid additional confusion. The numeric range is between 0.01 and 0.99 “to reflect the uncertain nature of intelligence estimates” [28]. The probability estimation should be complemented by a confidence level, although no set of confidence expressions was mentioned in the original publication [27].

- The North Atlantic Treaty Organization (NATO) formulated a NBLP scheme in 2016 as part of the NATO Allied Joint Doctrine for Intelligence Procedures 2.1 [29]. The original publication is confidential, but the scheme is described in articles like [28]. With a set of five expressions, the NATO scheme is the smallest one that we could find.

- The Professional Head of Intelligence Assessment (PHIA) Probability Yardstick was developed in the early 2000 by the UK Defence Intelligence. The yardstick has multiple versions and we refer to the latest version that was published in a technical report in 2019 [30, p. 29]. The yardstick’s goals are to foster standardized processes that promote objective results over time. The yardstick does not cover the full probability range, but leaves 5% gaps between categories. The gaps create a distance between the categories and shall motivate experts to confidently decide on one probability range. Another feature is that probabilities can also be expressed in fractions which allows to address the lower and upper bound of categories (e.g., the fraction \( \frac{1}{20} \) denotes the upper bound of “remote change” (5%)).

- The established term words of estimative probabilities (WEP) originates from a proposal of Sherman Kent in 1964 [31] which was, however, never adopted by the Central Intelligence Agency (CIA). The WEP scheme is considered the first of its kind. Instead of translating expressions to numeric thresholds or ranges, expressions are translated to precise numbers with tolerances (e.g., “Probable” translates to 75\% ± 12\%). The WEP scheme has three probability gaps.

While the use of NBLP schemes are certainly an easy and structured way to facilitate uncertainty communication, a remaining problem is that users must agree and commit to use the proposed VPEs for specific probability values. Recent research, however, indicates that “most people struggle to suppress the meanings they normally associate with such terms” [28]. The overall goal is to prepare and motivate people to express their estimations numerically. Instead of establishing a standard VPE set for all, we see the development of individual NBLP schemes as part of the solution. A human-agent team can engage in a co-learning process to develop an individual NBLP scheme where the model can be trained as secondary task and the human is familiarized with expressing uncertainty numerically over time. The concept is illustrated in the next section.

3. Applying co-learning to manage uncertain communication

3.1. Use case: smartphone repair

To illustrate the co-learning process, we adopt the fictional use case from Van Zoelen et al. [32] in which a service technician collaborates with an AI agent to analyze a broken
smartphone and to perform the necessary steps to repair the smartphone. In the collaboration scenario, the AI agent supports the service technician with relevant information that it (pro-actively or re-actively) retrieves from the company’s knowledge base. On the other hand, the service technician provides feedback on the results to the AI agent. The collaboration model of the service technician and the AI agent qualifies as hybrid team which “consists of agents. An agent is an entity that is autonomous, intentional, social, reactive, and proactive [33]. So a human is an agent. A machine can also be an agent, but only if it meets the criteria above” [8]. Additionally, the scenario addresses the challenges of developing a taxonomy model and of providing an appropriate communication model which were identified by Van den Bosch et al. [8].

If we consider a specific scenario where a damaged smartphone irregularly shuts down, the service technician might make several observations of the behavior of the smartphone during the analysis that must be communicated to the AI agent with a level of uncertainty due to the inconsistent behavior of the smartphone. As the service technician practices a think-aloud approach while performing the analysis as his primary task, the formulation of probability estimations is secondary. In that case, we argue that it is more beneficial to permit the service technician to begin by annotating his assumptions and observations with VPEs, until reasoning on the knowledge base can provide precise numeric probability estimates. To achieve a mutual understanding of uttered VPEs, the hybrid team must engage in a co-learning process.

3.2. Phases of the co-learning process

To prepare the service technician for expressing uncertainty in numeric values only, we consider three phases of communication in the hybrid team. In the first phase, the service technician uses only VPEs to convey probabilistic observations to the AI agent. In this scenario, the service technician is introduced to the NATO scheme for familiarizing himself with the thoughtful application of VPEs. After the service technician is familiarized, the service technician can add his own expressions. After each expressed observation or belief of the service technician is recorded, the AI agent confirms each utterance by repeating it.

In the second phase, the service technician continues to use VPEs. After a statement is recorded, the AI agent replaces the VPE by the calculated numeric estimate when repeating the statement. In case the service technician is not satisfied with the suggested probability, the service technician can correct the suggested probability by proposing a numeric value which the AI agent records for future processing.

In the third phase, the service technician only uses numeric values to define the probabilistic character of an event. As soon as the service technician feels comfortable using only numeric values, the service technician’s beliefs are less prone to be misunderstood by (human) colleagues and shared beliefs can be easier processed. The three phases are summarized in Table 2.

In all phases, whenever the agent is asked by the service technician to query the knowledge base, the agent tries to mitigate subjective bias by presenting recorded knowledge about probabilistic events. If there are shared beliefs, the annotated probability will always be based on numeric aggregation. However, based on the service technician’s preferences, the numeric range or value can still be translated to a VPE, following the service technician’s individual and subjective interpretation.
4. Discussion

4.1. Application of VPEs

In general, researchers argue to use numeric values to express probabilities, because numeric values are more precise and unambiguous [5, 28, 34]. However, practitioners may still prefer to communicate estimations using VPEs (e.g., in intelligence analysis [28]), which makes it inevitable for researchers in the domain of applied science to find means to work with VPEs. NBLP schemes seem to be an easy way to achieve that. Main criticisms of established NBLP schemes are that they are based on the opinions of a small group, were not developed with evidence-based methods [10, 35], and lack empirical validation [5]. Evidence-based NBLP schemes were presented by Ho et al. [35] and Wintle et al. [36]. Also, the use of VPEs might be insufficiently expressive in scenarios where a high granularity of probability values is required.

Apart from the expression of pure probabilities, some biases must be managed when creating estimations via VPEs. Friedman and Zeckhauser [37] argue that experts should express their level of confidence together with their probabilistic assessment. Confidence addresses the robustness of the probabilistic assessment information that may be acquired in the future. The confidence level should not be confused with confidence of statistical analysis [25]. Several NBLP schemes like the IPCC scheme are used in combination with a confidence scale. Recent findings, however, indicate that experts and non-experts conflate their estimated probability with their confidence level [38]. Moreover, native and non-native speakers seem to have different numeric interpretations [39, 40]. Another important aspect is the aggregation of multiple probability estimations. Participants aggregate VPEs by guesswork instead of mental computation [41] which can lead to incorrect conclusions. For instance, “when the advisors both say an event is ‘likely’, participants will say that it is ‘very likely’” [42] instead of increasing their confidence that occurrence of the event is in fact just “likely”.

Nonetheless, there is an ease of use by applying VPEs, and “in most everyday situations, verbal probabilities may not only suffice, but may also be ideal, given that they afford free-flowing communication.” [34, p. 11]. It is important to keep in mind that there

---

**Table 2.** Outline of the three phases of the co-learning process in the context of the use case.

<table>
<thead>
<tr>
<th>Phase I</th>
<th>Phase II</th>
<th>Phase III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service technician uses solely VPEs. AI agent uses only VPEs if requested by service technician or probability cannot be determined from collected data.</td>
<td>Service technician uses VPEs and optionally numeric values. AI agent uses exclusively numeric probabilities.</td>
<td>Service technician and AI agent use only numeric values to communicate probabilities.</td>
</tr>
</tbody>
</table>
will always be biases when people communicate uncertainty. On the one hand, VPEs are used to "save face", to maintain credibility as an expert, to avoid being blamed, and just to be polite in some cases [5,34]. On the other hand, the application of numeric probability expressions can also be biased [43,44].

We argue that a co-learning process within hybrid teams can significantly contribute to reduce aforementioned biases.

4.2. Modelling and processing VPEs

An established view is that numeric translations of VPEs are not crisp boundaries, but can be modeled by using fuzzy sets [3,45,46,47]. While the application of fuzzy sets is agreed on, both triangular and trapezoidal fuzzy set membership functions are used in practice [18]. Several authors [46,48,49] proposed probabilistic pragmatics models that are based on the Rational Speech Act model [50,51] to formalize communication with VPEs and therefore validate and optimize the choice of VPEs used between speaker and listener. While this approach seems promising, it must be stressed that rational acts depend on the current information state and can turn out sub-optimal in hindsight after new information was acquired. Therefore, it is imperative to apply (1) appropriate methods to establish co-learning processes for estimating uncertainty and (2) to identify and capture the relevant context when realizing a co-learning process due to the highly context-dependent nature of VPEs. The lottery of guessing the next ball in a urn (used by Herbstritt and Franke [48]) might be too simple to represent complex scenarios. Also, lotteries are criticized to be too time-intensive [52]. While our described use case illustrates a potential realization of the co-learning process to process VPEs, empirical research is needed to validate these claims. Hence, we formulate the following two research questions to be addressed in future research:

RQ1 Which algorithms, mechanisms, and methods are most appropriate to establish co-learning processes for estimating uncertainty in the context of collaborative tasks in human-agent teams?

RQ2 How to appropriately identify and capture the task context to respond to the highly context-dependent nature of verbal probability expressions?

5. Conclusion

If possible, probabilities of uncertain states or events should always be conveyed using numeric means. In consideration of the preference paradox and depending on the context, the use of verbal probability expressions (VPEs) might be (temporarily) more beneficial or even inevitable. We propose a co-learning process within hybrid teams to diminsh the preference paradox and other biases. In this process, human and AI team members are familiarized with the numeric translation of VPEs until numeric probability values can be efficiently communicated by both parties. Future work will address the research questions which were mentioned in section 4.
References

LLM-Augmented Agent-Based Modelling for Social Simulations: Challenges and Opportunities

Önder GÜRCAN
Norwegian Research Center AS (NORCE), Universitetsveien 19, Kristiansand, Norway
e-mail: ongu@norceresearch.no

Abstract. As large language models (LLMs) continue to make significant strides, their better integration into agent-based simulations offers a transformational potential for understanding complex social systems. However, such integration is not trivial and poses numerous challenges. Based on this observation, in this paper, we explore architectures and methods to systematically develop LLM-augmented social simulations and discuss potential research directions in this field. We conclude that integrating LLMs with agent-based simulations offers a powerful toolset for researchers and scientists, allowing for more nuanced, realistic, and comprehensive models of complex systems and human behaviours.

Keywords. Large Language Models (LLMs), Agent-Based Simulations, Social Systems Modeling, Research Directions

1. Introduction

Large Language Models (LLMs) have experienced a rapid expansion in adoption across a multitude of research domains and practical applications in recently. This swift proliferation is largely attributed to their remarkable ability in understanding, generating, and translating human language with unprecedented accuracy and fluency. Industries ranging from healthcare [1,2], where they assist in patient care and medical research, to finance [3], for analyzing market trends and automating customer service, have leveraged the capabilities of LLMs to enhance efficiency and innovation. Furthermore, in the realm of academia, these models are instrumental in analyzing vast datasets for insights [4], thereby accelerating research outcomes in fields such as social sciences, linguistics, and computer science. The versatility and evolving sophistication of LLMs have thus cemented their role as a cornerstone technology that is reshaping the landscape of both industry and research, fostering new methodologies and approaches across disciplines.

In social sciences, one area where LLMs can be effectively applied is social simulations [5]. Social simulations are used to model and analyze the complex interactions within social systems, including factors like individual behaviours,
group dynamics, social norms, and institutional structures. These simulations aim to understand, predict, or examine hypothetical scenarios within social systems. In that sense, a technique used for social simulations is agent-based modelling (ABM). ABM is used to design and simulate the actions and interactions of autonomous agents (individuals, entities, or organizations) to assess their effects on the system as a whole. ABM can utilize AI techniques to enhance the models’ complexity, adaptability, and realism. So far, ABM has already practically incorporated AI techniques like Machine Learning (ML) [6,7,8,9], Reinforcement Learning [10,11] and Inverse Reinforcement Learning [12] in various social simulation studies. However, the potential of LLMs in assisting in understanding of complex social systems has yet to be realized. There are few studies focusing on such incorporation recently. Even though these studies provide some practical solutions, they lack a well-defined conceptual baseline which is necessary to explore general-purpose architectures and methodologies – eventually extending existing ones – that are effective to seamlessly and systematically integrate social simulations and LLMs.

Furthermore, LLMs can streamline and augment various other aspects of the ABM process, such as literature reviewing, data preparation and interpretation, calibration of parameters, sensitivity analyses, analyses of results. We have already begun to see some impacts of LLMs on these activities. But they are scarce and do not provide a clear vision in terms of social simulations. Drawing upon these observations, in this paper, we outline prospective avenues for research into the augmentation of agent-based simulations through the integration of LLMs. We articulate a structured exploration of the symbiotic potential between LLMs and ABM frameworks, aiming to advance the methodological foundations and enhance the analytical capabilities of social simulations.

The remainder of the paper is organized as follows. We first provide a background on LLMs and ABM for social simulations in Section 2. After, we identify the conceptual baseline by evaluating various methodologies for building multi-agent systems in Section 3. Then, we provide an overview of some main research directions useful to develop the idea in Section 4 and finally in Section 5 we conclude the paper.

2. Background

2.1. Large-Language Models (LLMs)

A large language model (LLM) can be defined as a function that is finding, considering a series of tokens (such as words, word fragments, punctuation, emojis, etc.), which tokens are most probable to follow next. As of today, there are numerous LLMs accessible to the public. Google’s Bard, based on the more efficient and compact PaLM 2 LLM with 340 billion parameters, is accessible for free via a web browser\(^1\) and API, featuring a diverse training dataset [13]. Meta’s LLaMA\(^2\), aimed at AI research advancement, offers various models with up to

\(^1\)https://bard.google.com/chat, last access on 07/02/2024.
\(^2\)https://llama.meta.com, last access on 07/02/2024.
70 billion parameters available to researchers through application [14]. OpenAI’s GPT models utilize transformer architecture for dynamic output generation [15], with various versions accessible differently: GPT-3.5 is free via a web interface³, while GPT-4 requires a subscription, with API usage also being pay-per-use based on tokenization for natural language processing. LLM APIs are interfaces that allow developers to access the capabilities of these advanced LLMs in their applications. APIs can be integrated into applications using any programming language that can make HTTP requests, typically by sending the prompt for the LLM to process along with various parameters that adjust the LLM’s behavior (such as the version of the language model to use and the temperature that controls the randomness of the generated output).

An LLM can be viewed as a non-deterministic simulator with the ability to role-play an endless array of characters. Essentially, LLMs can be fine-tuned by exposing them to specific roles so that they can simulate human-like interactions. Fine-tuning is achieved by training the model on a curated dataset that embodies the language, knowledge, and nuances of the roles it is expected to perform. This process involves adjusting the model’s parameters so it better aligns with the patterns, vocabulary, and decision-making processes characteristic of the target roles. During fine-tuning, the model learns to prioritize responses that reflect the specific traits, expertise, or persona of the roles in question. This is often done by using a smaller, more specialized dataset after the model has been pre-trained on a broad corpus of text, allowing it to adapt its vast general knowledge to more narrowly defined contexts and behaviors. However, such dataset as well can be relatively large.

A well-known architecture for building systems that leverage LLMs on specialized relatively large datasets is called Retrieval Augmented Generation (RAG) [16]. In this architecture, the dataset is broken into chunks (such as a couple of paragraphs or a page) and then those chunks are sent to an LLM and they are turned into a vector. Each chunk will have a vector (i.e. a series of numbers) which is a numeric representation of the essence of that chunk. After, each time right after a prompt is sent, its vector is calculated as well using the same LLM. Then, the closest chunks are found by performing a mathematical comparison between the chunk vectors and the prompt vector. Finally, these chunks are used as part of the prompt.

### 2.2. ABM in Social Simulations

ABM stands as a pivotal technique in the exploration and understanding of social systems through computational simulations [17]. Social systems, in this context, refer to complex networks of interactions among individuals, institutions, and their environments (which can include factors like individual behaviours, group dynamics, social norms, and institutional structures). The multifaceted nature of social systems allows individuals to play multiple roles simultaneously (such as a parent, employee, consumer, and citizen in a human society) each with its own set of expectations and norms.
ABM for social systems aims to mimic social processes by simulating the actions and interactions of agents, which represent individuals or entities within these systems, to predict and understand complex phenomena. In other words, ABM enhances the study of social systems by offering a bottom-up modeling approach, where the macro-level phenomena of interest emerge from the micro-level interactions of agents. This capability is invaluable in social science, where understanding the emergence of complex social phenomena from simple rules of interaction can provide profound insights into the nature of social order, the evolution of norms and institutions, and the dynamics of social change. Social simulations provide a powerful means to examine hypothetical scenarios, test theories of social behavior and interaction, and explore the potential effects of policy decisions without the ethical and practical constraints of real-world experimentation.

However, analyzing the results of ABM to understand social systems presents several challenges, primarily due to the complexity and dynamism inherent in both the models and the systems they seek to represent. Firstly, ABMs often generate vast amounts of data through the simulation of interactions among numerous agents over time, making it difficult to discern clear patterns or draw straightforward conclusions. The emergent phenomena, a hallmark of ABM, while valuable for understanding the macro-level outcomes of micro-level behaviors, can complicate analysis as these outcomes are not always predictable or linear. Secondly, interpreting the results of ABMs in a manner that is meaningful for policymaking or theoretical advancement requires bridging the gap between complex, often technical, model outputs and the conceptual frameworks of social sciences. This often demands interdisciplinary collaboration to ensure that the insights generated are both scientifically rigorous and socially relevant.

Besides, in addition to ABM, social scientists employ various other methods to study social systems. Surveys and questionnaires are used to collect large-scale data on individual attitudes, behaviours, and experiences. Interviews and ethnography offer in-depth qualitative insights into social phenomena, capturing the nuances of human behaviour and social interactions. Case studies provide detailed analysis of specific instances or events, allowing for a deep understanding of complex social processes. Experimental and quasi-experimental designs are utilized to establish causal relationships between variables. Statistical analysis and computational techniques, including network analysis and data mining, are applied to analyze and model large datasets. Lastly, content and discourse analysis are employed to examine communication patterns and the construction of meaning within social contexts. Some of these methods are already used in ABM studies.

3. The Conceptual Baseline for LLM-Augmented Social Simulations

Integrating LLMs into ABMs for social simulations offers a transformational potential for understanding complex social systems. However, such integration requires a conceptual baseline to connect both domains. A conceptual baseline is a clear and coherent foundational framework that outlines key concepts, variables, assumptions, and relationships within a given system or model. It serves as a reference point for understanding the dynamics and behaviour of the system being
studied, and it is subsequently used for model development and analysis. Such a conceptual baseline can be established using the existing engineering methodologies dedicated to multi-agent systems (MAS). A MAS is typically viewed from four primary angles: agent, interaction, environment and organization.

**Agent-oriented methodologies** prioritize individual agents, their autonomy, internal states, and decision-making [18,19,20]. They use the Beliefs, Desires, and Intentions (BDI) model for agent cognition and define agents’ capabilities and roles. This approach enables detailed modeling of complex, autonomous agents that make decisions based on perceptions and fulfill roles in multi-agent interactions.

**Interaction-oriented methodologies** focus on agent communication and coordination dynamics, using protocols and messages to define interaction patterns [21,22,23]. Protocols ensure structured information exchanges, while negotiation and coordination mechanisms support joint decision-making and action synchronization. Social norms and conventions establish a framework for predictable interactions and adherence to shared conduct rules.

**Environment-oriented methodologies** emphasize the environment’s role in agent interactions, focusing on shared resources, stigmergy for indirect coordination, and affordances that dictate agent actions based on their capabilities [24,25]. **Organizational-oriented methodologies** abstract the MAS regarding groups, teams, and broader organizational structures [26,27,28,29,30,31,32]. These methodologies focus on how agents are arranged and interact within larger entities, defining roles, responsibilities, and relationships through groups and teams. Organizational structures, such as hierarchies or networks, dictate the flow of information and control among agents. Policies and regulations set out the rules governing behaviour within the organization, guiding agent actions and interactions.

The organizational-oriented MAS approach is particularly well-suited for modeling social systems and integrating LLMs due to its emphasis on structured interactions and roles within a complex system. This approach mirrors the hierarchical and networked nature of social systems, where entities (agents) assume specific roles and responsibilities that are governed by established norms and policies. Such an organizational structure enables the clear definition of roles for LLMs, facilitating their integration as agents with specific functions related to language understanding, generation, and processing. This not only enhances the system’s ability to mimic human social structures but also leverages LLMs’ capabilities in natural language processing to improve communication and coordination among agents. By aligning the structural and functional aspects of MAS with the inherent properties of social systems and the strengths of LLMs, the organizational-oriented approach offers a robust framework for capturing the complexity and dynamics of social interactions, making it a superior choice for these applications. Therefore, we claim that an efficient ABM tool augmented by LLM should support the organization-oriented conceptual baseline.

In that light, we propose defining agents in social simulations as social agents that are role-playing one or several predefined characters [33,34,35]. A social agent’s skills are the role playing capacities they have through their interactions with environments, and are used in a community of other role players who also inhabit in these environments.
4. Research Directions

In this section, we explore key research paths that seem pertinent for transforming social simulations with the help of LLMs.

4.1. Literature Reviews

The volume of scientific literature is overwhelming [36,37], with varying levels of quality across publications. There is a necessity for tools that can search, evaluate, and summarize scientific literature both objectively and efficiently. LLMs can significantly address the challenges of the traditional review process through their advanced capabilities in processing vast amounts of text data efficiently [37,38,39] and they are less likely to cherry-pick the literature to support desired hypotheses (i.e. reduced researcher bias) [40]. By automating the initial screening and summarization of literature, LLMs can help researchers navigate the issue of information overload, enabling them to quickly identify relevant studies without compromising the breadth or depth of the review. Their ability to analyze and summarize texts in multiple languages can also overcome language barriers, providing access to a wider range of literature.

4.2. Modeling Architectures

As discussed in Section 3, the organization-oriented approach which is defining agents as role playing actors provides a solid grounding for modeling social agents. However, there are several organization-oriented architectures out there and the fit for the LLM-augmented social simulations is yet to be studied. Moreover, research should also be done for the effective design and re-use of roles of social agents. Effective design of roles is necessary for obtaining effective insights (see Section 4.5) and effective re-use of roles is necessary for large and repeatable social simulations. Furthermore, LLMs can be used to generate agent-based models and scenarios using natural language and related qualitative data.

4.3. Data Preparation

The data collection for social simulations is time consuming and expensive since it faces several key challenges, including capturing the complexity of social systems, ensuring high-quality and relevant data, addressing ethical and privacy concerns, integrating diverse data sources, and achieving accurate modeling and simulation. Additionally, the calibration and validation of models, managing computational constraints, fostering interdisciplinary collaboration, and adapting to dynamic social systems add layers of complexity. Finally, ensuring the generalizability and transferability of the models to different contexts or populations is a significant challenge, requiring a careful, methodical approach and often interdisciplinary collaboration. LLMs have the potential to significantly enhance the data collection process for social simulations by offering solutions to many of the challenges outlined. For capturing the complexity of social systems, LLMs can process and analyze large volumes of text data from diverse sources [41,42], providing a nuanced understanding of social dynamics that can inform more accurate and com-
prehensive models. Their advanced natural language processing capabilities enable the integration of varied data types, from structured data to unstructured text, facilitating the creation of richer, multidimensional datasets.

4.4. Datafication

Datafication is the transformation of complex social interactions and phenomena into quantifiable data, allowing for real-time tracking and predictive analysis [43]. LLM-augmented social agents can play a pivotal role in the process of datafication. This data can then be analyzed to uncover patterns, trends, and insights about social behaviors and systems. For instance, social agents interact within simulations, they continuously generate data that can reflect changes over time, including how social systems evolve in response to external pressures or internal dynamics. This dynamic aspect of data generation is crucial for studying processes of social change, innovation diffusion, and the emergence of social norms.

4.5. Obtaining Insights

Social simulations often produce data that are too voluminous and too complex to curate and analyze. It is shown that using LLMs it is possible to get insights from data [44]. Hence, insights can be obtained social agents by simply entering into dialogues with them. If prepared properly, it is possible to simulate a synthetic population of a vast array of social agents representing human experiences and perspectives [45], which can provide a more precise depiction of human behavior and social dynamics than what is achievable through traditional methods [5]. With proper conditioning [46], social agents can role-play characters that have beliefs and intentions and that provide accurate and objective answers to users’ questions [47]. They are ideal obtaining insights as they “can rapidly answer hundreds of questions without fatigue” and “need fewer incentives than humans to give reliable responses” [48]. Recent studies show that LLMs can have the ability to make judgements quite well aligned with human judgements [48]. The information collected from various dialogues can then be organized as quantifiable data that can be statistically analyzed for providing insights into broader social trends and patterns. Such an approach can then be used in generating hypotheses and validating them in human societies [46,35].

4.6. Explainability

LLM-augmented social agents can generate natural language explanations for their actions, decisions, and the underlying mechanics of the simulation [49,50]. This can make the behavior of these agents more understandable to researchers (from various disciplines, not just those with computational backgrounds), stakeholders, and the general public, translating complex algorithms and decision-making processes into easily digestible explanations. Moreover, by leveraging the vast knowledge and understanding of social dynamics embedded within LLMs, these augmented agents can provide contextual insights into their behavior. For instance, they can explain how certain social norms, historical events, or cultural aspects influence their actions within the simulation, offering a deeper understanding of the modeled social phenomena.
4.7. Platforms and Tools

Effective LLM-augmented social simulations require sophisticated tool support, tailored to manage the complexities of dynamic social interactions and datafication processes. These tools should facilitate the seamless integration of LLMs with simulation platforms, offering features for easy configuration, real-time adjustments, and ethical data handling. Importantly, the design of these tools must be grounded in an organization-oriented conceptual baseline (see Section 3), ensuring they align with the structured dynamics of social systems and support focused, interdisciplinary research. This approach enhances the accessibility, accuracy, and ethical compliance of simulations, enabling researchers to deeply explore and understand social phenomena.

5. Conclusions

LLMs offer a transformative framework for the simulation and analysis of social systems. In this study, we proposed different LLM interventions that span the entire social simulation pipeline. In addition to reasoning the behavior of involved agents, LLMs can enable more prosperous and intuitive interactions between users and simulated agents. This can make social simulation more accessible and user-friendly, allowing individuals from various disciplines, including social sciences, healthcare, urban planning, and environmental studies, to utilize them. Such democratization can enable these professionals to model and analyze complex systems relevant to their fields without needing deep expertise in computer science. As a result, we can expect an increased collaboration between computer scientists and experts from other fields, leading to more interdisciplinary approaches and innovative solutions to complex real-world problems.

The potential advantages of Large Language Models (LLMs) merit serious consideration. However, it is essential for scientists and developers working with LLM-enhanced Agent-Based Modeling (ABM) tools to also entertain the possibility that these tools could, under certain circumstances, hinder rather than advance scientific knowledge. This means that while LLMs offer significant epistemic benefits, they also pose epistemic risks if scientists rely on them as partners in producing knowledge [51]. Because, treating LLM-augmented ABM tools as collaborators in scientific research exposes scientists to the risk of falling into illusions of understanding, which is a class of metacognitive error that occurs when individuals have mistaken beliefs about the depth or accuracy of their own comprehension [52,53].

Acknowledgements

The work reported here is part of the URBANE project, which has received funding from the European Union’s Horizon Europe Innovation Action under grant agreement No. 101069782.
References


[38] Zhang T, Ladzhak F, Dumus E, Liang P, McKeown K, Hashimoto TB. Benchmarking...


Investigating Labeler Bias in Face Annotation for Machine Learning

Luke HALIBURTON\textsuperscript{a,b,1}, Sinksar GHEBREMEDHIN\textsuperscript{a}, Robin WELSCH\textsuperscript{c}, Albrecht SCHMIDT\textsuperscript{a}, and Sven MAYER\textsuperscript{a}

\textsuperscript{a}LMU Munich, Germany
\textsuperscript{b}Munich Center for Machine Learning (MCML), Germany
\textsuperscript{c}Aalto University, Finland

ORCID ID: Luke Haliburton https://orcid.org/0000-0002-5654-2453, Sinksar Ghebremedhin https://orcid.org/0000-0002-2874-2909, Robin Welsch https://orcid.org/0000-0002-7255-7890, Albrecht Schmidt https://orcid.org/0000-0003-3890-1990, Sven Mayer https://orcid.org/0000-0001-5462-8782

Abstract. In a world increasingly reliant on artificial intelligence, it is more important than ever to consider the ethical implications of artificial intelligence. One key under-explored challenge is labeler bias — bias introduced by individuals who label datasets — which can create inherently biased datasets for training and subsequently lead to inaccurate or unfair decisions in healthcare, employment, education, and law enforcement. Hence, we conducted a study (\(N=98\)) to investigate and measure the existence of labeler bias using images of people from different ethnicities and sexes in a labeling task. Our results show that participants hold stereotypes that influence their decision-making process and that labeler demographics impact assigned labels. We also discuss how labeler bias influences datasets and, subsequently, the models trained on them. Overall, a high degree of transparency must be maintained throughout the entire artificial intelligence training process to identify and correct biases in the data as early as possible.

Keywords. bias, machine learning, crowdworkers, annotation, labeler bias

1. Introduction

Artificial intelligence (AI) is rapidly becoming involved in numerous areas of life, making far-reaching decisions such as granting loans and hiring people. Amazon analyzes customers’ purchasing behavior\textsuperscript{2}, Netflix studies entertainment preferences\textsuperscript{3}, and Facebook uses social interactions to tailor content to their users\textsuperscript{1}. Data collection, processing, and prediction are key pillars of AI applications. Although AI is a powerful tool, the fundamental reliance on data can be problematic due to the potential for bias to be em-

\textsuperscript{1}Corresponding Author: Luke Haliburton, luke.haliburton@ifi.lmu.de.
\textsuperscript{2}https://www.gigaspaces.com/blog/amazon-found-every-100ms-of-latency-cost-them-1-in-sales
bedded in datasets, creating unintended consequences. One under-investigated contributing factor to biased AI tools is labeler bias, which results from cognitive biases [2] in crowd workers and other dynamics in the labeling process [3]. Many AI applications rely on crowdsourcing platforms to label their data, yet they usually do not consider whether they are utilizing a diverse population of labelers [4]. A biased labeler pool could lead to unfair outcomes for certain groups, such as women, ethnic minorities, or people from disadvantaged neighborhoods. Therefore, it is crucial to examine labeler pools with a critical lens to avoid bias and create a more fair and transparent process.

Investigating labeler bias is essential to understand how labelers influence datasets, but existing studies in this area are limited in scope. Recent work has demonstrated that rater identity plays a significant role in labeling toxicity for online comments [5, 6]. One critical paper by Bender et al. [7] sheds light on how human biases can be unintentionally perpetuated in machine learning, highlighting that biases introduced in the labeling stage can propagate through to end decisions. In response to this issue, researchers in machine learning have attempted to model and correct for bias effects [8, 9, 10]. In general, bias can be partially attributed to stereotypes, which occur when one assigns traits to an individual based on preconceived notions about their group [11]. The Stereotype Content Model (SCM) is an established practical theory explaining stereotypes, such as perceived warmth and competence, that has been applied in Human-Computer Interaction (HCI) (e.g., [12, 13]). However, there is a lack of work applying the SCM to characterize biases introduced into datasets by crowdsourced labelers.

In this paper, we address the gap in the existing literature by investigating stereotypes and bias in labeling tasks. We conducted a survey (N=98) asking crowd workers to label a series of human faces from the FairFace dataset [14]. We selected faces with equal representation from seven ethnicities and two sexes and recruited labelers with the same balanced demographic distribution. We asked labelers to rate the portraits based on income and perceived warmth, competence, status, and competition. In this way, we investigate relationships between stereotype perceptions and income within and between demographic groups. We developed this study design to answer the following research questions:

**RQ1** What is the impact of stereotypes held by labelers on their estimates of others in labeling tasks?

**RQ2** What is the impact of the ethnicity or sex of labelers on their estimates of others in labeling tasks?

Our results show that labelers possess stereotypes independent of their own demographics and that these stereotypes impact the labels they assign. Further, we show that the ethnicities of both labelers and portraits impact the predictions assigned by the labelers. Our results indicate that labeler bias is a function of both labeler demographics and characteristics of the labeled subject, suggesting that recruiting a diverse labeler pool may not be enough to counteract the bias. Overall, this paper contributes evidence for the existence of labeler bias and discusses its consequences. In particular, we contribute an investigation of bias in the context of a face-labeling task using a publicly available...
dataset. Our findings raise awareness for labeler bias, which we hope leads to researchers and practitioners critically examining and revisiting current practices in data labeling.

2. Related Work

In this section, we first present prior work on data bias in machine learning, examining historical bias and labeler bias. Next, we introduce the stereotype content model (SCM) and how it relates to computer science and our work.

2.1. Bias in Machine Learning

Biases can be introduced in machine learning models and algorithms at multiple stages. Algorithms can contain systematic biases embedded by the moral concepts of developers [15, 16]. However, data bias is a more prevalent concern in intelligent systems. As Mueller [17] describes, the quality of a system is coupled with “the quality of the data provided, following the old slogan garbage in, garbage out.” It follows that an AI system will make biased decisions if it is trained on biased data. Notable sources of bias, which we detail below, include historical bias, non-representative sampling, and labeler bias.

2.1.1. Historical and Sampling Bias

Historical bias occurs when a system is trained on data resulting from real, biased scenarios. In an early example, St. George’s University Medical School introduced a new computer system that systematically denied admission to women and people with ‘foreign-sounding names’ based on historical data [18, 19]. Similar issues persist in many modern systems. A CV screening tool at Amazon preferentially hired men even after explicit references to gender, race, and sexual orientation were removed from the dataset [20]. Algorithms for crime prediction typically rely on historical crime data in which ethnic minorities are over-represented5. Even using online proxies to remove sensitive characteristics continues to lead to biased decisions [19, 21]. Such systems can have a drastic impact on the lives of real people. For example, minority groups have an increased likelihood of being stopped and searched by the police, based only on immutable characteristics which they cannot control [22].

Generating datasets based on emerging data can also lead to biased results. Bender et al. [7] describe the experience of collecting data through the internet: “white supremacist and misogynistic, ageist, etc. views are over-represented in the training data, not only exceeding their prevalence in the general population but also setting up models trained on these datasets to amplify biases and harms further.” Women are also underrepresented on platforms used for data collection [23], which results in decisions based on male-skewed data sources. Although issues of historical bias and non-representative sampling continue to be an issue, they are not the focus of this paper. Rather, we aim to investigate the under-explored potential impact of bias within the people labeling data to generate biased datasets.

Sampling bias has also been identified in face annotation tasks. Da Silva and Pedrini [24] found that an emotion classifier trained on one cultural group was inaccurate when

---

used on a different cultural group. Scheuerman et al. [25] recommend embedding race and gender information into databases and classifications to increase transparency.

2.1.2. Labeler Bias

Labeler bias occurs when individuals annotate a dataset and embed bias into the resulting data. This often occurs unintentionally, as Wall et al. [26] have argued that unconscious biases can influence judgments and lead to inaccurate conclusions in visual analytic feedback tasks. In one poignant example of labeler bias, prior work has shown that people from Western cultures tend to rate people from other cultures as being less attractive than themselves [27]. Consequently, an attractiveness dataset labeled primarily by people from Western cultures risks having this bias embedded. Any system that uses such a dataset as a basis for decisions would subsequently perpetuate the bias of the labelers. Past work in machine learning has shown that fairness and accuracy can be aligned, motivating a push towards more fairness in training as a simultaneous push towards increased accuracy [28]. Prior research has found that even highly experienced labelers fail to produce unbiased labels [29].

Perhaps most relevant to our study is recent work in CSCW by Goyal et al. [5] on rater identity. They found that rater identity (i.e., African American, LGBTQ, or neither) significantly influenced how raters annotated toxicity in online comments. They suggest that raters who self-identify with the identities targeted in online comments provide additional nuance and more inclusivity in trained models. Similar work found that other social variables (e.g., conservatism) also impact toxicity labels [6]. Past work has identified that labeler bias also stems from socio-economic contexts and the application of power structures within annotation companies [3]. In response, researchers have proposed more fair and human-centric crowdsourcing frameworks considering demographics and appropriate compensation [30]. In the presented examples, we can see that labeler bias can stem from a number of characteristics of the labeling group. In this paper, we extend the body of work on labeler bias by exploring whether demographic clusters of labelers can have an impact on labeling task estimates.

Past work has attempted to find solutions to account for labeler bias. Several strategies to model labeler bias include using knowledge about the ground truth [8], bayesian methods [9], or multi-task Gaussian Processes [10]. Geva et al. [31] recommend that labelers for testing and training datasets be distinct groups since they found that subjective NLP labels produced by a group do not generalize well. Instructions for annotators have also been found to embed bias [32, 33]. In the context of face annotation, Engelmann et al. [34] argue that ‘secondary’ (i.e., subjective) characteristics may not be appropriate attributes for facial recognition systems to predict.

2.2. Stereotype Content Model

The Stereotype Content Model (SCM) is a theory that explains how people develop stereotypes about other people. The SCM proposes that people primarily use two dimensions, warmth and competence, to assess other people. These dimensions prognosticate emotional prejudices, which in turn lead to discrimination [11, 35, 36]. People group themselves based on what they perceive to be similarities between themselves and oth-

\[\text{Attractiveness Test: } \text{https://attractivenesstest.com/}\]
ers. Across-group prejudice is a judgment on an emotional, cognitive, or behavioral level towards another group [37].

The warmth-competence model is a well-researched [38, 39, 40, 12, 41] fundamental theory in social psychology with broad implications for both social perception and social interaction [42]. The model helps to explain how we form stereotypes about different groups of people, positing that we judge groups based on how warm and competent we perceive them to be [43, 44]. Warmth is the evaluation criterion by which a person is perceived as friendly and trustworthy, while competence shows how capable and compelling a person is perceived to be. Generally, our reaction to individuals from certain groups is governed by how warm and competent we perceive them to be. For example, people generally see elderly individuals as warm but not competent and therefore react with pity. Groups seen as competent but not warm are met with envy, while those perceived as both competent and warm tend to be admired [36]. The SCM has also been extended to include competition and status, which are particularly influenced by how an individual views a specific group relative to their own group [36].

Nicolas et al. [45] made a significant contribution by developing dictionaries for stereotypical content. These dictionaries simplify the study of stereotypes and speed up the identification of social biases in AI, social media, and other textual sources. Subsequent work has used these dictionaries to automatically identify the stereotypical language in news coverage [46] or to mitigate stereotypical language through anti-stereotypes [47]. The SCM has also been used to understand human-machine interactions. For example, McKee et al. [48] used the SCM to explore how individuals react to different digital avatars. They found that users increasingly perceive the system as being warm if it appears in the role of an assistant and cold if it appears in the role of a competitor. The SCM has also been used in HCI to investigate stereotypes in personas [13] and determine the social acceptability of mobile devices [12]. In our work, we are investigating the way that labeler stereotypes influence the labels they assign during annotation tasks. As such, we use the SCM, a well-established model, to understand how stereotypes vary across our labeler population.

3. Method

We conducted an online survey where participants from various demographic backgrounds labeled portraits of varying ethnicities and sexes based on income and the SCM. For this, we balanced the participants’ self-reported ethnicity.

Since ethnicity labels are not clearly defined [49], we aligned our ethnicity categorization with the FairFace dataset. The following seven groups listed in FairFace will be referred to as ethnicities7 in this work: (1) Black, (2) East Asian, (3) Indian, (4) Latino, (5) Middle Eastern, (6) Southeast Asian, and (7) White.

3.1. Dataset Preprocessing and Portrait Selection

For our study, we selected portraits from the FairFace dataset (License CC BY 4.0) [14]. This dataset consists of images of people and was specifically developed to be balanced

---

7We use the term “ethnicity,” as it encompasses more social aspects and is a broader term than “race,” although they are often used interchangeably in practice. We also note that ethnic distinctions are non-specific, but such labels are commonly used in Machine Learning applications such as the FairFace dataset.
in ethnicity, sex, and age. The authors generated this dataset to foster the development of fair and inclusive machine learning models. We selected the FairFace dataset because it provides us with a high probability of finding high-quality images across a wide range of demographic categories. The ethnicity, gender, and age tags for the images were labeled by Amazon Mechanical Turk (MTurk) users based on a two-thirds majority vote [14]. The majority vote labeling process is common practice for labeling [50], but consequently, we do not have a ground truth for the labels. However, to our knowledge, this is the most extensive dataset available with diverse ethnicities and sexes and was, therefore, the most suitable choice for this study.

We defined several criteria to filter the images in the dataset and create a subset for our study. Our research questions concern ethnicity and sex, so we designated age as a control variable and only selected images in the dataset within the range of 27 to 40 years old (the narrowest age filter provided by the dataset). To avoid any confounding factors, we also defined the following criteria: 1) Only one visible, camera-facing person with a neutral facial expression, 2) A neutral background with no cropped edges, 3) In color (i.e., no black & white images), and 4) No glasses/sunglasses/headgear.

As the FairFace dataset contains more than 100,000 images, we could not manually filter all images by our predefined criteria. Therefore, we created a preprocessing script with the following functions: (1) We removed any images with age labels lower than 27 or higher than 40 (2) We detected faces within each image using the python library “face-recognition” and removed any images where the number of faces was equal to zero or greater than one. (3) We detected face landmarks using the python library mediapipe [51] and used these landmarks (see Figure 1) to remove images where the subject is not facing the camera. A face is considered to be non-frontal if it deviates on the x-axis (see Figure 1b) or the y-axis (see Figure 1c) beyond a threshold of 0.09, or if both thresholds are crossed, indicating that the face is completely non-frontal (see Figure 1d). If none of the thresholds are crossed, the face is considered to be frontally aligned (see Figure 1a). (4) We detected facial expressions using a machine learning model based on9. We selected only images with neutral facial expressions. (5) Finally, we were left with 1,834 images which we manually filtered. Three authors evaluated each of the remaining images and only selected those that fit all of the above criteria. The three authors triple-coded and only included images where all agreed. This resulted in 56 portraits, four for each sex and ethnicity combination.

8https://github.com/ageitgey/face_recognition
9https://github.com/priya-dwivedi/face_and_emotion_detection
3.2. Participants

We recruited 98 participants (49 female and 49 male) from Prolific\textsuperscript{10}. Participants were between 18 and 52 years old ($M = 26.1, SD = 6.9$). The participants were equally distributed among the seven ethnicity categories from the FairFace dataset, and age. They live in 21 countries\textsuperscript{11}. Participants were compensated at a rate of 10€ per hour for a total of 3€. The study was approved by the ethics committee within the LMU Munich University Faculty. Participants’ income, as indicated in £ on Prolific, varied across ethnicities, shown in Figure 2. We analyzed the correlation between participant income and their income estimates for the portraits to prevent estimation bias. A Pearson correlation showed no significant correlation ($r = -0.025, p = .403$), so we assume that participants’ own income does not influence the results.

3.3. Study Procedure

We used Prolific, a crowdsourcing marketplace, to gather data as it can provide demographic information about the participants. We created a separate posting for each sex and ethnicity category and used built-in demographic filtering features to ensure that we recruited an equal number of participants from each category.

The participants were first given a brief overview before providing informed consent and completing a demographic questionnaire. The demographic information was also provided by Prolific, but we collected it in the survey to ensure that the ethnicity and sex labels matched the terminology used in the dataset. We then asked participants to respond to questions associated with portraits of people. Each participant was presented with one randomly chosen portrait from each ethnicity and sex category, resulting in 14 portraits per participant.

For each portrait, the participants responded to SCM questions about perceptions of warmth, competence, status, and competition based on [36]\textsuperscript{11}. Several sub-scale questions are averaged to score ratings of warmth, competence, status, and competition on

\textsuperscript{10}Prolific: https://www.prolific.co

\textsuperscript{11}Full study information is provided at https://github.com/mimuc/labeler-bias.
Figure 3. Warmth-Competence ratings displayed from the perspective of the portraits and the labelers, including a 95% confidence interval. All ratings are clustered near neutral (3) for both warmth and competence. 

4. Results

To investigate the relationship between demographics and labels, we performed two-way ANOVA models (Type III, $\alpha = .05$) using Greenhouse-Geisser correction [53] where the sphericity assumption is violated. Note that although it would be possible to compare all levels on all factors and their interaction with post hoc tests, we refrain from doing so. First, test-corrections will be very conservative for pairwise comparisons due to the high number of levels. Second, we were only interested in showing that our factors can explain variation on the dependent variable. As such, we do not analyze which specific biases are present in specific groups.

4.1. The Impact of Stereotypes on Estimations (RQ1)

In line with the SCM [36], Figure 3 shows the Warmth-Competence ratings assigned to the portraits by the labelers. Figure 3a shows the stereotypes assigned to each PORTRAIT$\text{Ethnicity}$ while Figure 3b presents the stereotypes assigned by each LABELER$\text{Ethnicity}$. All ratings are clustered near neutral warmth and neutral competence.
Table 1. The Pearson correlations for each of the stereotype variables and the estimated income. We also computed Linear mixed models that take into account the nested structure in the data. However, showed no noteworthy difference from the simple correlations, so we only report the Pearson correlations for brevity.

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warmth</td>
<td>.093</td>
<td>.045</td>
</tr>
<tr>
<td>Competence</td>
<td>&lt;.001</td>
<td>.541</td>
</tr>
<tr>
<td>Status</td>
<td>&lt;.001</td>
<td>.773</td>
</tr>
<tr>
<td>Competition</td>
<td>&lt;.001</td>
<td>.431</td>
</tr>
</tbody>
</table>

Figure 4. Correlation between mean status and income. Each subplot represents a portrait ethnicity and the points in each plot show how labelers of each ethnicity rated the portraits.

We conducted a Pearson correlation analysis for warmth, competence, status, and competition against the estimated income. Table 1 shows that three out of four stereotype variables have a significant positive correlation with income. Competence, status, and competition all covary significantly with income, with status being the most positively correlated. Only warmth did not significantly covary with the income estimates.

In general, the biases were in the predicted directions [36]. For example, participants generally estimated a low income for a person they rated as low status, and vice versa. Figure 4 illustrates this phenomenon, showing a positive correlation between Status and Income.

4.2. The Impact of Demographics on Estimations (RQ2)

Our second research question asks whether participant demographics impact their perceptions of stereotypes and their income estimations. Figure 5 shows the income labels assigned to each PORTRAIT_Ethnicity by each LABELER_Ethnicity. We conducted an ANOVA using the interaction effect of LABELER_Ethnicity and PORTRAIT_Ethnicity. The results, shown in Table 2, reveal that the LABELER_Ethnicity
Figure 5. Estimated income as a function of LABELEREthnicity and PORTRAITEthnicity. Grey borders indicate the cases where LABELEREthnicity and PORTRAITEthnicity match.

Table 2. The two-way ANOVA results for the Income estimates and the four stereotype variables for LABELEREthnicity and PORTRAITEthnicity.

<table>
<thead>
<tr>
<th>LABELEREthnicity</th>
<th>PORTRAITEthnicity</th>
<th>L × P</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>df</td>
<td>F</td>
</tr>
<tr>
<td>Income</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Competence</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Status</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Competition</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

The interaction effect is significant for all variables except warmth. For an exemplary chart that shows income as a function of LABELEREthnicity and PORTRAITEthnicity, see Figure 5.

5. Discussion

In this work, we set out to answer our two research questions. In the following, we discuss how our results address the research questions, the consequences for machine learning practice, and the limitations of our study.

5.1. Labelers Exhibit Bias

Our results provide evidence that labeler bias exists in two ways. First, our findings indicate that labelers hold stereotypes about the people depicted in the portraits. 

Figure 5. Estimated income as a function of LABELEREthnicity and PORTRAITEthnicity. Grey borders indicate the cases where LABELEREthnicity and PORTRAITEthnicity match.

Table 2. The two-way ANOVA results for the Income estimates and the four stereotype variables for LABELEREthnicity and PORTRAITEthnicity.
they are tasked with annotating (see Figure 3a) and that these stereotypes have an impact on their subsequent income labels (see Table 1). Second, Table 2 demonstrates that LabelerEthnicity and PortraitEthnicity have a significant impact on perceptions of stereotypes and income labels. Our results show that LabelerEthnicity significantly impacts status, PortraitEthnicity significantly impacts all variables, and the interaction effect is significant for all variables except warmth.

We found that income estimates were correlated with stereotype perceptions (RQ1) and that both stereotype perceptions and income estimates were impacted by LabelerEthnicity and the PortraitEthnicity (RQ2). Interestingly, Sex only had a significant effect on warmth for this task. This is in line with psychological research showing how stereotype judgments are made across cultures and, in particular, how gender stereotypes influence perception. In this domain, where labelers are annotating portraits of humans, we can conclude that labeler bias exists, depends on labeler demographics, and can be explained using stereotype content. These results are in line with recent findings in CSCW by Goyal et al. demonstrating that toxicity labels for online content are influenced by labeler self-identification.

5.2. Implications for Machine Learning

Our results shed some light on the impact of the human-aspect of machine learning. We have shown that labels vary with labeler demographics for annotation tasks involving portraits. This is important because prior work has shown that the majority of workers on MTurk are from the United States and India, and Levay et al. found that over 70% of MTurkers identify as white. Consequently, the status quo of gathering labels on crowdsourcing platforms without considering who is doing the labeling should be reconsidered, as it will almost certainly lead to a non-balanced labeler pool and subsequently biased labels. However, as shown in Figure 3a and Table 2, stereotypes occur in labeling tasks involving images of people independent of labeler demographics. Recruiting labelers from a wide variety of populations should result in labels that are biased in a way that is consistent with societal biases, but the bias will still exist. Therefore, it remains an open research question as to whether it is possible to combat labeler bias through recruiting or to correct it post hoc. Post hoc methods have been proposed for bias in machine learning labels (e.g., [8]), but typically require knowledge of ground truth. Ground truths for social issues are complicated by the fact that representation in data is connected to the unequal distribution of power in society. On such social issues, where ground truths may be fuzzy or non-existent, and society may be biased against particular groups, there is space for crucial future research to develop solutions. Past work by Miceli et al. suggests expanding data documentation and incorporating social contexts, which could be an important step toward ensuring fair, transparent data pipelines.

---

Note that we did not test all comparisons post hoc as this procedure would not be informative to our study for two main reasons: First, one would need to apply very conservative $\alpha$-level corrections in order to avoid Type-II error inflation. Second, we were not interested in characterizing the specific bias of a certain group but rather intend to show that variation within the estimates can be explained by considering the interaction of labeler demographics and portrait demographics.
5.3. Limitations & Future Work

The most notable limitation of our study is that we have only explored one specific labeling task, namely annotating secondary characteristics of faces. This labeling scenario was chosen because it was likely to highlight the impact of stereotypes while still being rooted in a realistic scenario, such as making hiring decisions, which have been shown to be impacted by stereotypes [60]. However, further work is required to understand how stereotypes and labeler demographics impact more abstract labeling tasks, such as image segmentation or product labeling. There are many high-stakes scenarios relevant to specific tasks across the field of machine learning that may be impacted by this phenomenon. Since we have now shown that there is a bias in this scenario, we call on future work to explore additional scenarios that are relevant to specific tasks in machine learning. For example, labeler bias may be relevant in detecting and classifying objects in autonomous driving tasks. Quantifying how these results generalize to other tasks is crucial to understanding when labeler demographics and their interactions with content must be accounted for.

Another limitation derives from the methodology used to create the FairFace dataset. The ethnicity, sex, and age labels in the dataset were created using a majority vote method on Amazon Mechanical Turk. The labels, therefore, are not necessarily a ground truth. Similarly, there is no ground truth for the income estimates since this information was not included in the FairFace dataset. Despite these limitations, this is the best available dataset we could locate with a balanced representation and labeled demographics and was, therefore, the best option available for this work. Future studies should investigate the magnitude of bias as a difference from the ground truth. Resolving this issue is not trivial, as it likely requires that a new database of images be generated with associated income levels provided by the image subjects, but it would be a worthy endeavor to further investigate and characterize this information.

6. Conclusion

In this paper, we investigated the existence and impact of labeler bias. We recruited 98 participants to engage in an online survey where we asked them to estimate the income and rate the perceived warmth, competence, status, and competition of people from multiple ethnicities and sexes portrayed in a series of images. We selected the portraits from the FairFace dataset using predefined exclusion criteria to create a balanced set of portraits. We found a significant relationship between income estimates and perceptions of competence, status, and competition. Additionally, the results indicate that the main- and interaction-effects of LABELEREthnicity and PORTRAITEthnicity significantly affect estimations, while LABELERSex and PORTRAITSex did not significantly impact the results. This insight poses a major challenge for AI applications, as it implies that datasets annotated by a non-diverse set of labelers are likely to carry stereotypes. Thus, we recommend that further research develops guidelines for responsible dataset generation and that researchers and practitioners reconsider the status quo for data labeling.
References


Human-AI Collaboration in Smart Manufacturing: Key Concepts and Framework for Design

Maria HARTIKAINEN1, Guna SPURAVA, and Kaisa VÄÄNÄNEN
Tampere University, Computing Sciences

ORCiD ID: Maria Hartikainen https://orcid.org/0000-0001-5997-0368
ORCiD ID: Guna Spurava https://orcid.org/0000-0001-5323-9806
ORCiD ID: Kaisa Väänänen https://orcid.org/0000-0002-3565-6021

Abstract. Demographical reasons and the increasing demand for improved production efficiency are steering the transformation within the manufacturing domain towards smart manufacturing. This entails introducing artificial intelligence (AI), data analytics, and automation to improve the efficiency, productivity, and flexibility of manufacturing processes. With the integration of AI, there is a shift from humans merely interacting with technology to actively collaborating with it, especially with AI-enabled agents. This shift brings changes in work practices and tasks. Hence, comprehensive understanding of the phenomenon becomes central for the design of human-AI collaboration that genuinely contributes to effective production and supports operators’ well-being. This scoping review study aims to shed light to the evolving landscape of human-AI collaboration in smart manufacturing by presenting six key concepts derived from an analysis of 23 academic papers. Based on the findings, we propose a framework that offers an initial basis for the design of human-AI collaborative systems for smart manufacturing.

Keywords. Artificial Intelligence, Human-Centred AI, Human-AI Collaboration, Human-Machine Interaction, Smart Manufacturing, Human-Centric Smart Manufacturing

1. Introduction

The increasing demand for improved production efficiency as well as demands stemming from demographical issues, like aging workforces, are steering the transformation within the manufacturing domain towards Smart Manufacturing [13][18]. Smart manufacturing, similarly to Industry 4.0, refers to the use of smart technologies like artificial intelligence (AI), data analytics, and automation to improve the efficiency, productivity, and flexibility of manufacturing processes [9][16][18]. Continuation to smart manufacturing is Human-Centric Smart Manufacturing (HSM) that similarly to EU-led Industry 5.0 emphasises individual well-being and addresses social challenges, placing human factors, particularly operator well-being, at its core [2][4][6][13][20][23]. With the integration of AI, there is a shift from humans merely interacting with

1 Corresponding Author: Maria Hartikainen, maria.hartikainen@tuni.fi.
technology to actively collaborating with it, especially with AI-enabled agents [7][15][24]. Enabled by AI, technology takes proactive role, as machines now perform tasks autonomously, and the roles between the human and machine are becoming more interchangeable [7][17]. Hence, the machine actors, referred as AI agents, are increasingly viewed as colleagues or teammates, signifying a new era of artificial colleague where human and AI collaborate towards common goal - human-AI collaboration [7][15][22][24]. This new form of interaction characterised by interdependencies and teamwork means changes to the work tasks and processes, especially for the operators, the workers in the manufacturing domain [2][15][23]. Hence, comprehensive understanding for the design of human-AI collaboration in smart manufacturing that genuinely is effective and supports operator’s well-being as the operators are the crucial part of the successful implementation of the collaboration [7][13][19][23]. Because of the collaboration with AI-enabled agents and integration of AI in work life, there is a need for human-centred artificial intelligence (HCAI), aiming to prioritise human needs over technology [3][12]. HCAI seeks to provide efficient solutions and positive outcomes for users and society, augmenting human abilities rather than replacing them [11][20][28]. Ethical development and use of AI are integral to HCAI, ensuring fairness, trustworthiness [10][11][22][27], whereas explainable and transparent interaction between humans and AI enhances trust and user understanding in human-AI collaboration [8][12][26]. Additionally, autonomy and human control must be carefully considered in human-AI collaboration [12][25]. Despite the topicality, the human-centred approach to the collaboration design and the operator’s well-being in changing work environments is under-explored. Hence, in this study, we review the related literature to understand, how do support the operator well-being by design and how to acknowledge the human factors in the collaboration design. We approach the study aim with the following research question: What are the main concepts related to human-AI collaboration that have been suggested by existing literature?

In this article, we present a scoping literature review that identifies the key concepts to acknowledge and address in the human-AI collaboration design [14]. Building upon these insights, our study proposes an initial human-AI collaboration framework for smart manufacturing, providing a foundation for designing and implementing effective and efficient collaboration strategies that support operator’s well-being.

2. Method

2.1. The scoping review

We approached our research objectives with a scoping literature review to understand the existing knowledge on the topic, as well as to identify the key concepts associated to the topic in the related literature [14]. First, we developed a review question to reflect the study aim. Second, we defined the inclusion/exclusion criteria. We included articles that are in manufacturing domain and are at least partly concentrating on human-AI collaboration. We included peer-reviewed articles written in English that we had a full access. We excluded articles concentrating solely on human-robot collaboration, as robots normally have specific physical forms and movements that might affect the human’s perception and expectations related to them. However, we did include articles that had robots as one example of AI-enabled technologies. Also, we excluded papers that were technical, related to the enabling AI technologies or programming of those, as
your aim was in understanding the human factors. Third step was to search literature relevant to our scope. We utilised Google Scholar because it covers all the biggest databases. We searched with different compositions of key words describing humans and AI collaborating or working together like “human-AI collaboration / cooperation” and “human-AI AND manufactur*”. This search returned 452 articles that we screened based on the title and abstract using our inclusion / exclusion criteria. We included 32 articles to full-text screening. We read these papers and based on our inclusion/exclusion criteria we included 15 papers in our study corpus. After this, we used snowballing method, and based on the full text screening, we included eight more articles. Hence, total 23 articles were included to our study corpus. Figure 1 shows the review process. Detailed list of the reviewed articles is in Appendix 1.

2.2. Analysis

We conducted reflexive thematic analysis to the data withdrawn from the reviewed papers with a purpose of find overarching concepts and their meanings in relation to the studied phenomena in review literature [5]. We began by familiarising ourselves with the data by reading the papers in our study corpus and the first author coded all the data utilising Atlas.ti -tool [1]. Even though our approach was reflexive analysis, we had four predetermined codes related to the key features of HCAI, because we wanted to understand how these are acknowledge in the related literature. These codes were: transparency, explainability, trust, and control. However, as our study aim is to understand the phenomena and we adapted reflexive analysis method, we left space for additional codes that rise from the analysed literature. Hence, during the first round of coding we combined deductive and inductive coding, and we added codes as they arose from the reviewed literature. After the preliminary coding, the authors discussed about the codes, they refined, combined, and separated them as needed. This led to 23 codes that formed the codebook. Our codebook is presented in table 1. Next, we conducted a second round of coding using the codebook. After the second round of coding, we conducted thematic analysis independently to identify recurrent concepts. We discussed about our individual analysis and combined our findings. Eventually, we defined six key concepts associated to the studied phenomena in the literature.

3. Findings

We identified six key concepts related to the human-AI collaboration from the reviewed literature. Table 1 presents the codes, the concepts and their description, and the literature included to each concept. In the text, we are referring to article with [A#]. Detailed list of referred articles is in Appendix 1.
### 3.1. Collaboration goal

Collaboration goals refer to the objectives set for the collaborative actions performed by humans and AI. These goals are aimed at achieving specific outcomes that leverage the strengths of both humans and machines to improve productivity, efficiency, quality, safety, and innovation in the manufacturing process [A3, A19, A20]. In manufacturing, collaboration goals can be short-term goals (e.g. solving a problem) or long-term goals (e.g. enhanced productivity), and these goals may change over time [A11]. Goals should be well defined at the beginning of the collaboration design, so that informed design decisions can be based on these goals. In addition, expected benefits should be defined before the design. In the manufacturing domain, human-AI collaboration offers various benefits, and it seamlessly integrates into every phase of the product construction process, from design to predictive maintenance [A19, A20]. It provides an effective solution for overcoming limitations in real manufacturing environments [A2]. Moreover, it enhances the production process and creates more versatile and engaging job opportunities for employees [A11, A16, A18]. The measurement and metrics for the collaboration success can be planned based on the goals and expected benefits. These measurements can be quantitative (e.g. product efficiency) or qualitative (e.g. operator’s well-being). Once the collaboration has been introduced and running, based on the defined measurements, human-AI collaboration performance can be measured, and adjustment can be done to the collaboration design as needed [A5].

### 3.2. Collaboration team

Collaboration team is formed with two or more agents, human and machine actors, that work together pursuing common goals [A9, A10, A15, A18]. In manufacturing, these agents can be humans, AI-agents, or AI-enabled robots [A17, A18, A23]. All agents have individual capabilities, strengths, and weaknesses [A8, A15]. Agents’ strengths and weaknesses should be mapped and understood beforehand, so that collaboration tasks can be planned based on these for effective team performance [A8, A11]. From human-centred perspective, AI strengths should be used to enhance and augment the operator’s skills [A4, A11, A12, A15, A20]. In smart manufacturing, the strengths that AI can pose are physical strength, speed, scalability, repeatability, and quantitative capabilities for complex analytical approach or data processing, and this way they can reduce the
operator’s mental workload [A1, A2, A16, A18, A20]. In addition, AI can help capture the equipment conditions and prevent system failures proactively. However, several manufacturing operations remain manual, as humans can perform these better and cannot be replaced by AI systems [A2]. Humans pose superior skills in unpredictable physical work, teamwork, social interactions, applying expertise, creativity, and managing [A2, A16, A18]. In addition, humans tend to be better to address anomalous situations and provide flexible solutions in case of need as well as evaluate the outcome of decision [A16, A18]. Moreover, in many cases regardless the strengths, humans will still be expected to be in control [A22, A23].

3.3. Skills

All agents possess individual skills, or capabilities, that can be acquired and developed through collaboration, training, and learning. In smart manufacturing, collaborating with AI and the fast-evolving technologies and changes is work processes, demand of new skills is high [A15, A16, A17, A19, A21]. Establishing appropriate skills represent the decisive factor to keep up not only with current but also future technology as the dynamic nature of the system underlines the constant evolution of skills among all actors involved [A7, A23, A17]. Perceptual, cognitive, emotional, and motoric skill demands on the agents are determined in the design [A11]. If these demands are higher that the capacity of an individual, there can be negative impacts on system performance, trust in AI, and worker well-being [A7, A11, A21, A19]. In addition, the current techno-centred design of smart manufacturing systems tends to demand extreme skills from the human operators as they are expected to handle any unexpected situations efficiently [A16]. At the same time, operators need to be able to operate and manage these adaptable AI systems, which requires corresponding skills [A7, A11]. Hence, training is a cornerstone of smart industry [A15, A21, A19]. Implementing a new system that heavily involves human-AI collaboration could demand significant training for workers to adapt to the new system, and there might be resistance to change [A21, A17]. Training is important not only because of the well-being of the operator, but also because inexperienced or under-trained personnel are prone to committing human errors [A2, A23]. Especially because in manufacturing, AI is increasingly applied in use cases with potentially severe consequences for humans [A7]. In addition, learning is an important part of collaboration, and it is important for the skill development of all agents. In human-AI collaboration in smart manufacturing, all agents possess ability to learn from each other or from the collaboration - this way collaboration team benefits mutually [A4, A12, A15, A19]. The dynamic nature of the overall system brings requirements where the capabilities and skills of all the actors change over time, hence, continuous learning is necessary [A4, A16, A11, A19].

3.4. Collaboration task

Collaboration goals are met with one or more collaborating tasks that are conducted by the collaborating team. Tasks are formed with activities that are conducted by the agents individually or collaboratively [A15, A18]. Starting from the analysis of team and task structure, the skills of the team actors are identified and linked to the different teaming activities [A9, A12, A15]. Tasks are carried out via interactions with agents and collaborative systems [A11, A13, A17]. Tasks have well-defined structure, and they have allocations [A12, A13]. An important design decision is the distribution of tasks between
the agents [A11]. Each agents’ capabilities should be key divider of the tasks, so that the
task allocation supports operator’s improved ergonomics, safety, and well-being by
eliminating or reducing monotonous, hazardous, and physically demanding tasks
traditionally performed by humans [A2, A16, A20, A22]. Tasks should be formed so that
AI compensates operator’s shortcomings or limitations in cognitive (e.g. complex data
analysis), physical (heavy or dangerous jobs), or even sensorial capabilities (e.g.
operator’s state-of-mind) to deliver the best of both worlds, and to explore human
augmentation to create manufacturing work that is more productive [A4, A11, A12, A15,
A18, A20]. By handling physically and mentally demanding tasks, AI agents ease the
burden on workers and that way promote the operator’s well-being, safety, and
ergonomics [A11, A20, A22, A23]. From the human-centred perspective, there is a need
to develop task allocation and teamwork in human-AI collaboration teams so that human
workers feel they are in the loop and that human remain meaningful and manageable
[A11]. Task allocation must be revised resiliently to adapt to the changes in the dynamic
manufacturing environment [A11, A13].

3.5. Collaboration system

Collaboration tasks and most of the interactions between the agents are carried out via
collaboration system. Hence, a key element for successful human-AI collaboration is a
careful design of the coordinating system involved [A9]. In smart manufacturing, these
systems are usually socio-technical systems called Cyber-Physical-Systems (CPS) [A6,
A11, A13, A16, A17]. CPSs comprises of humans, AI and the physical system that are
in connection with the surrounding physical world and its processes [A11, A16, A17].
These systems should promote communication and understanding between the agents,
as well as situational awareness, by offering possibility and interface to share knowledge
about each other and their roles in the current task or in the collaboration process, or by
showing the location of the other team member or predicting their next action in the
collaboration process [A1, A6, A11, A14, A18]. In addition, collaborating system should
aim to improve and promote learning of the agents and they should be dynamic and
resilient, so that they can answer changing needs and dynamic environments in
manufacturing, and the design should support this [A1, A11, A15, A16, A17]. In smart
manufacturing, human-AI collaboration is moving beyond traditional interaction
mechanisms, as smart technologies can allow humans to convey information with AI
systems through multiple channels by integrating advanced human-machine interfaces
that offer information related to the context and the situation that is relevant to the
interaction [A6, A14, A18, A20]. With AI’s help, there are more possibilities for the
interaction, so the preferred interaction ways should be defined based on the skills and
preferences of the human, as well as to fit to the context they are going to be used [A18,
A20]. In addition, integrating multiple modalities is an effective approach to overcoming
limitations in a real manufacturing environment [A2].

3.6. Communication

Communication is a central element of human-centeredness within human-AI
collaboration [A7, A19]. Human-AI communication is dynamic, contingent upon the
messages being exchanged at a specific moment within a particular context [A12, A14].
Communication between the agents serves two important purposes: (i) to convey
essential procedural information for progressing the manufacturing operations, and (ii)
to provide feedback that rewards or encourages good collaborations between humans and AI-agents, leading to mutually positive human-AI relationships [A3, A12, A19]. Human-AI collaboration team can rely on a mixture of verbal-based and nonverbal signs for natural and intuitive communication to coordinate team behaviours and learn from each other [A4, A20]. Important part of communication is understanding. Understanding between the collaborating agents promotes safety and efficiency in the collaboration [A7, A12, A18]. In human-centric smart manufacturing, human intent understanding is important, as well as appropriate level of expectations [A18], hence the machines need to possess some level of “empathy” as they need to actively collaborate with humans based on dynamic human intent and align with human’s aspirations and motivations [A12]. This promotes understanding between the two agents [A11, A18]. From human-centred perspective, to confirm and reinforce the role of the human, explainability is required to human to understand the usage and results of AI and AI systems should provide easy-to-understand explanations of its actions and recommendations [A3, A20]. Transparency is needed in order the human to trust machine decisions - people need to know how an AI system derives its conclusions and makes its actions [A20]. Also, transparency is important to set appropriate expectations for both sides. From the human-centred perspective, it is important to define what information to convey, via what channel, and in what level, as communication and mutual understanding is important part of successful human-AI communication and trust building [A1, A3, A9, A12, A18].

4. Discussion

We identified six key concepts associated to the human-AI collaboration from the reviewed literature. All the concepts need careful considerations, and they should be acknowledged in the human-AI collaboration design for the collaboration to be genuinely effective. Based on the study findings, we propose an initial framework for human-AI collaboration in smart manufacturing. It presents the key concepts to acknowledge in the human-AI collaboration design, their relationships, and dependencies.

![Proposed framework to present the key concepts of human-AI collaboration in smart manufacturing.](image)

Figure 2. Proposed framework to present the key concepts of human-AI collaboration in smart manufacturing.

The integration of AI into manufacturing is a socio-technical process, influenced by not only technological possibilities but also by social factors such as operators’ well-being, trust building, and human interpretation of AI systems that should be acknowledged in the collaboration design. Utilisation of AI makes AI-related factors, like explainability and transparency, relevant in this paradigm. There is a need to investigate how should they be acknowledged in the collaboration design. In the reviewed literature it is clear, that implementing AI to the work practises and processes may radically change the
traditional work tasks and practises, and this of course requires very different skills than the more traditional manufacturing tasks. The current techno-centred design of smart manufacturing systems tends to demand extreme skills from the human operators as they are expected to handle any unexpected situations efficiently. As the mismatch between the skills and skills requirements strongly affects to the operator’s well-being, it is crucial to ensure the appropriate skills of operator with training and learning to promotes the well-being of the operator. Hence, instructions for mapping the operator’s skill requirements or training needs are currently lacking. Existing studies are mostly literature reviews, and empirical user studies are needed to understand the human side in the collaboration. In addition, the operators’ point of view is under-explored in current literature, even though operators are the crucial part of the successful implementation of the collaboration [13][23]. Hence, for future research, we suggest the following topics:

- Human-AI collaboration design guidelines should be developed, to support the design of effective human-AI collaboration in manufacturing companies. The concepts identified in this paper, and the proposed framework, can serve as a basis for these guidelines.
- Methods to map operator skills and skills demand, and appropriate training schemes should be developed.
- HCAI related factors such as transparency, ethics, explainability, should be studied in relation to the human-AI collaboration in smart manufacturing.
- Design approaches and methods are needed for the dynamic nature of the overall system where the skills of all the actors change over time.

Limitations of our study are that even though we covered several different terms related to human-AI collaboration, in the snowballing phase we comprehended that this topic is covered in papers related industry 4.0 and industry 5.0. In addition, we used ‘manufacturing’ as a keyword, however, we later learned that ‘production’ is often used as a synonym for it. Hence, we might not have covered all the articles related to the topic in our review. However, we were not aiming to systematic review, and in our opinion, the key concepts and their meaning we clear in the articles we reviewed. In addition, in the future research we can add to this knowledge by including data from more articles, if needed.

5. Conclusion

Human-AI collaboration in smart manufacturing fosters intuitive interaction between humans and AI, leveraging strengths for better system performance. Human-centeredness is crucial, with AI technologies promoting collaboration, augmenting human skills, and adapting to individual skills. By improving production processes and offering varied job opportunities, this collaboration enhances productivity and flexibility. This article contributes to the emerging field of human-AI collaboration in smart manufacturing by presenting key concepts associated to the topic. Based on the identified concepts we propose an initial human-AI collaboration for smart manufacturing framework that serves as a basis for researchers and industry professionals looking to delve into designing effective human-AI collaborations in smart manufacturing.
References

References – reviewed literature


A4 Bousdekis, Alexandros; Apostolou, Dimitriis; Mentzas, Gregoris (2020). A human cyber physical system framework for operator 4.0-artificial intelligence symbiosis. In Manufacturing letters, 2020

A5 Brauner, Philipp; Vervier, Luisa; Brilowski, Florian; Dammers, Hannah; Steuer-Dankert, Linda; Schneider, Sebastian; Baier, Ralph; Zieflé, Martin; Gries, Thomas; Leicht-Scholten, Carmen (2020). Organization Routines in Next Generation Manufacturing. Forecasting Next Generation Manufacturing: Digital Shadows, Human-Machine Collaboration, and Data-driven Business Models.


A9 Hoch, Thomas; Heinzl, Bernhard; Czech, Gerald; Khan, Maqbool; Waibel, Philipp; Bachhofner, Stefan; Kiesling, Elmar; Moser, Bernhard (2022). Teaming. AI: enabling human-AI teaming intelligence in manufacturing. Proceedings http://ceur-ws.org ISSN


Romero, David; Stahre, Johan; Wuest, Thorsten; Noran, Ovidiu; Bernus, Peter; Fast-Berglund, Åsa; Gorecky, Dominic (2016). Towards an operator 4.0 typology: a human-centric perspective on the fourth industrial revolution technologies. In Proceedings of the international conference on computers and industrial engineering (CIE46), Tianjin, China.


Trakadas, Panagiotis; Simoens, Pieter; Gkonis, Panagiotis; Sarakis, Lambros; Angelopoulou, Angelos; Ramallo-González, Alfonso P; Skarmeta, Antonio; Trochoutsos, Christos; Calvo, Daniel; Pariente, Tomas (2020). An artificial intelligence-based collaboration approach in industrial iot manufacturing: Key concepts, architectural extensions and potential applications. In Sensors 2020.

Tóth, Attila; Nagy, László; Kennedy, Roderick; Bohuž, Belej; Abonyi, János; Ruppert, Tamás (2023). The human-centric Industry 5.0 collaboration architecture. In MethodsX 2023.


Assessing the HI-ness of Virtual Heritage Applications with Knowledge Engineering

Delaram JAVDANI RIKHTEHGAR a,1, Ilaria TIDDI b, Shenghui WANG a, Stefan SCHLOBACH b, and Dirk HEYLEN a

a University of Twente, Enschede, The Netherlands
b Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

ORCiD ID: Delaram Javdani Rikhtehgar https://orcid.org/0009-0007-8266-8395, Ilaria Tiddi https://orcid.org/0000-0001-7116-9338, Shenghui Wang https://orcid.org/0000-0003-0583-6969, Stefan Schlobach https://orcid.org/0000-0002-3282-1597, Dirk Heylen https://orcid.org/0000-0003-4288-3334

Abstract. Virtual Heritage exhibitions aim to engage a diverse audience through the integration of Virtual Reality and various AI technologies, including Artificial Agents, and Knowledge Graphs. Understanding the nuances of human-agent interactions is crucial to fully harness the potential of these technologies and deliver personalized and captivating experiences. Evaluating the alignment of Virtual Heritage applications with the vision of Hybrid Intelligence – where humans and machines collaborate toward a common goal – presents a significant challenge. In this paper, we investigate the assessment of Hybrid Intelligence within the Virtual Heritage domain using Knowledge Engineering methods. Through the analysis of six different scenarios presented as workflows of tasks and input/output data, we identify and compare classical Knowledge Engineering tasks with HI-specific tasks to measure the level of HI-ness achieved. Our study focuses on evaluating the synergy achieved by mixed teams during various tasks as a measure of HI-ness. The findings provide insights into the effectiveness of Knowledge Engineering to identify HI aspects within existing applications, the potential for quantifying and improving HI-ness in an application, and the identification of modeling limitations.

Keywords. Virtual Heritage, Knowledge Engineering, Hybrid Intelligence, Human-Computer Interaction, Personalization

1. Introduction

We investigate the use of Knowledge Engineering methods to assess the nature of Hybrid Intelligence [1] in a virtual reality-based application within the domain of Cultural Heritage (CH). Recent advancements in Virtual Reality (VR) technologies have opened up new possibilities for attracting and engaging end-users through immersive and interactive experiences in the realm of art, history, and culture exploration [2,3]. To enhance accessibility to cultural heritage, institutions have integrated multimodal technologies to pro-

1Corresponding Author: Delaram Javdani Rikhtehgar, d.javdanirikhtehgar@utwente.nl
provide a more realistic and immersive visitor experience [4], attract diverse audiences [5], foster collaborative meaning-making [6], and elicit more natural interactions [7].

With the growing instances of Virtual Heritage (VH) [8] applications, virtual agents within intelligent environments play an important role in improving user experiences and satisfaction by facilitating effective communication and interactions. It is therefore crucial to understand the intricacies of human-agent interactions to leverage the full potential of AI technologies in VR and provide personalized and engaging experiences.

In pursuit of this understanding, we have designed an ontology that allows us to model complex scenarios involving heterogeneous data, models, agents, and interactions. This ontology facilitates the formalization of scenarios, such as (1) a virtual museum guide equipped with Theory of Mind capability (i.e., the cognitive ability to attribute and understand mental states such as beliefs, intentions, and emotions, enabling the interpretation and prediction of behavior based on these mental states [9]), assessing a visitor’s emotional states in real-time and adapting its painting description to increase engagement; (2) a collaborative and immersive exploration of the exhibition, where the virtual agent delivers a rich, educational experience based on user preferences, suggesting new paintings serendipitously and observing user reactions; and (3) a trustworthy museum tour where the user learns to trust the agent’s recommendations, as it provides transparent explanations on preferences, historical context, and the overall narrative flow of the exhibition.

The symbiotic collaboration of heterogeneous actors toward shared goals, the adaptability to each other’s capabilities, and the use of heterogeneous data and processing methods makes Virtual Heritage applications typical Hybrid Intelligence scenarios. The field of Hybrid Intelligence (HI) envisions the development of collaborative systems, where humans and intelligent machines operate in mixed teams collaboratively, synergistically, and proactively to achieve shared goals [1]. In a typical HI scenario, humans and artificial agents complement each other’s limitations — i.e. human stereotyping, error proneness, in-group favoritism, or short memory are mitigated by machine-driven decision-making, while human feedback ensures machine fairness, sample efficiency, and task generalizability.

Making a VR exhibition a hybrid intelligence application is crucial for maximizing collaboration between human visitors and AI-driven agents, particularly in the context where their collaboration is centered on enhancing the overall experience, facilitating exploration, and providing personalized tour. By leveraging the diverse strengths of both humans and machines, the exhibition can effectively address their common goal of engaging visitors and transferring knowledge. This collaborative approach amplifies the exhibition’s ability to tailor content delivery, guide immersive exploration, and maximize understanding and enjoyment of its content, ultimately fostering deeper engagement and satisfaction among visitors. In essence, integrating hybrid intelligence elevates the exhibition’s interactive capabilities and enriches the overall user experience.

In this context, a full understanding of the inner workings of the hybrid interaction between users and virtual agents in our scenario is still missing, making us lose the potential of a fully-fledged HI application. In other words, the challenge we face is to determine whether our scenarios are HI-compliant, and more specifically, how much HI-ness manifests in the Virtual Guided Tour (VGT) application. We focus on assessing the level of HI-ness in terms of synergy between humans and machines; the stronger the HI, the greater the synergy in collaboration.
To tackle this problem, this paper suggests the use of Knowledge Engineering [10] to measure the HI-ness of an existing application, focusing on the Virtual Guided Tour application. Knowledge Engineering allows to elicit, structure, formalize, and operationalize the information, knowledge, and tasks involved in knowledge-intensive applications. Methods such as CommonKADS [11] supported engineers in clarifying the structure of complex applications in the past. Its adaptation was recently used to define the most typical tasks, inputs/outputs, and knowledge roles (called the application’s Knowledge Model) in Hybrid Intelligence applications [12]. Here, we bring forward the idea that the HI Knowledge Model can be used as an analytical toolkit to measure the HI-ness of the VGT application. We call it the Hybrid Intelligence Knowledge Engineering toolkit (or HIKE), and aim at answering the following two research questions: (1) can we identify the HI aspects in our scenarios using HIKE, and if so, which ones? and (2) how can we use HIKE to evaluate the HI-ness of an application? In order to answer these questions, we start from a set of Virtual Guided Tour scenarios and then formally describe all processes, inputs and outputs involved therein to create a Knowledge Model for the VGT application. We then decompose the scenarios into workflows of tasks and inputs/outputs, and evaluate (i) the amount of specific HI tasks as compared to classical KE tasks and (ii) the strength of this HI-ness using the tasks’ qualitative analysis. The workflows ultimately help us assessing whether the Hybrid Intelligence Knowledge Engineering toolkit helps us sufficiently describe the multiple HI aspects of our scenarios.

2. Related Work

Our research is related to Virtual Heritage applications and Knowledge Engineering methods for Hybrid Intelligence.

Virtual Reality and Virtual Heritage Applications. The potential of extended reality (XR), notably augmented reality (AR) and virtual reality (VR), has been widely acknowledged by museums [13]. VR, in particular, has demonstrated its effectiveness in facilitating learning [14,15,16,17,18] and fostering student motivation [19]. In [20], a novel didactic evaluation method for Digital Cultural Heritage (DCH) learning in higher education, underscoring VR’s value in teaching, was introduced. The research highlights the experiential and engaging aspects of VR, emphasizing its compensatory role in guiding learners compared to traditional teaching methods.

VR has also proven instrumental in enhancing visitor experiences within museums [21,22]. [23] found that VR, by capturing users’ attention, strengthens engagement, leads to an improved overall user experience. The immersive nature of VR allows users to explore the realms of active imagination, creating a heightened sense of presence within the virtual environment [24]. [25] demonstrated that VR can stimulate the intention for physical visits to museums, emphasizing the profound impact of VR experiences. Authors of [26] emphasized the compelling opportunities VR provides for museum visitors to engage with places or objects that may be challenging to exhibit physically due to budget constraints, limited space, or staffing issues. The role of VR in creating interactive audience experiences, emphasizing how understanding user attention and behavior in VR informs the creative process, is discussed in [27].

The field of Virtual Heritage integrates various interactive technologies aimed at engaging a diverse audience and creating personalized visitor experiences [28]. One cate-
gory involves multimodal technologies which contribute to the immersive nature of VR experiences, thereby enhancing visitors’ overall experience [29]. Conversational agents have been incorporated within virtual museums to serve as guides to visitors and substantially improve the accessibility of information [30,31,32]. Furthermore, eye-tracking technology has emerged as a valuable tool in VR to offer insights into users’ behavior and interactions. This technology enables the creation of tailored experiences by monitoring and understanding visitors’ gaze patterns, allowing for the customization of content based on individual preferences [33,34]. Beyond behavioral analysis, eye-tracking is employed to discern visitors’ emotional states [35]. Additionally, eye-tracking technology has been leveraged to monitor visitor learning experience [36].

Knowledge Engineering and Hybrid Intelligence. Knowledge Engineering allows to design and formalize knowledge-based systems following software engineering methodologies [10]. Different modeling paradigms address different engineering aspects, i.e. CommonKADS [11] allows to describe models, MIKE [37] formalizes the execution of the models, and PROTEGE [38] allows collaborative knowledge acquisition and reasoning. CommonKADS has been applied to characterize a variety of single-agent scenarios (e.g. governance, smart grid management and robot control) [39,40,41], or multi-agent ones (e.g. supply chain management and traffic simulations) [42,43]. These scenarios are limited to the design of classical KE tasks, such as diagnosis, assessment, planning. New KE tasks that adapt to the open-endness of the modern applications have only been proposed in the Semantic Web and Hybrid Intelligence areas [12,44].

With the rise of Hybrid Intelligence as a field, several formalizations appeared. An official conceptualization of HI as a field (in the lines of the Human-Computer Interaction ontology [45]) is yet to be achieved [12]. The HI research agenda presented in [1] defines research challenges and solutions in HI (Theory of Mind for synergetic collaboration, reinforcement learning for adaptability to changing environments, societal and personal value-awareness integrated in the systems, knowledge graphs as background knowledge to increase trust in the team). A taxonomy to design HI systems is proposed in [46]. Standard models and design patterns for human-machine team collaboration were presented in [47,48,49]. A preliminary model to characterize and evaluating HI systems is presented in [50], based on a user-study that identifies the importance of team properties (boundedness, interdependence, competency, purposefulness etc.). An ontology describing actors, interactions and information processing types, with a set of common HI tasks and sub-tasks, is presented in [12]. Our goal is not to propose a new HI terminology, but rather apply the existing terminologies to assess the HI-ness of our application.

3. Preliminaries

In order to show how to use the method of [12] to measure the HI-ness of our Virtual Guided Tour application, this section introduces a few preliminaries, namely a set of relevant VGT scenarios and the basics of the HIKE toolkit.

3.1. Scenarios

In the VGT, a virtual agent accompanies a visitor (let us call her Sarah) through a VR exhibition. Information about the exhibition’s objects (e.g. paintings) is stored in a cultural
knowledge graph that the agent uses to guide Sarah throughout, recommending her relevant art objects and answering questions she may have. Sarah interacts with the agent, creating a personalized tour where her preferences and interest are considered.

**S1. Multimodal Interactions.** The agent captures Sarah’s gaze, facial expressions, and speech inputs, employing multimodal analysis and reasoning to understand her interests. Based on the interpreted multimodal inputs, the agent dynamically tailors its interaction (e.g. highlighting specific objects in the painting to guide visual attention or encourage hand grabbing, offering additional cultural background, recommending related paintings), thus providing Sarah with personalized information according to her interests.

**S2. Real-time Understanding of User States.** The agent employs Theory of Mind to assess Sarah’s beliefs, intentions, and emotions as she interacts with the agent within the virtual environment. The agent dynamically adjusts its responses to align with Sarah’s beliefs or intentions, provides emotional support to address her feelings, or changes its own beliefs, intentions, and emotions to improve future interactions.

**S3. Memory Integration for Personalization.** The agent, equipped with memory capabilities, observes Sarah’s actions, recalling her preferences (e.g. a favorite painter from a past session), and integrates them with information from previous users to enhance the exhibition experience. The agent adapts its explanations based on known user patterns and preferences, giving more emphasis to artists and styles resonating with users sharing similar tastes. The agent actively learns from its past actions, storing this knowledge in its memory, to inform and improve future interactions with users.

**S4. Perception Building.** An ongoing communication between Sarah and the agent takes place. The agent employs analytics, e.g. gaze-based measures, to learn information about Sarah’s interests or knowledge level. Simultaneously, Sarah gains insights into the agent’s capabilities and limitations, understanding the type of information it can and cannot provide, shaping her perception of the agent’s abilities.

**S5. Transparency and Explainability.** Intrigued by the agent’s highlighted paintings, Sarah asks for the reasons behind these choices. In response, the agent provides a transparent explanation, detailing the information it utilized (e.g. Sarah’s interest and preferences, her past interactions, and insights gathered from other users’ engagements) and the overall narrative flow of the exhibition. This transparent communication enhances Sarah’s trust in the agent’s guidance, enriching her overall experience.

**S6. Collaborative Pursuit of Goals.** Sarah and the agent engage in a collaborative exploration. The agent’s goal is to foster understanding of the exhibition’s theme by showcasing curated paintings, while Sarah focuses on learning different art styles. The agent adapts its guidance to accommodate both educational goal and artistic appreciation. Their collaborative efforts result in a comprehensive and immersive exploration, satisfying both individual and shared objectives.

### 3.2. HIKE – The Hybrid Intelligence Knowledge Engineering Toolkit

Taking inspiration from CommonKADS’ templates to describe the behavior of knowledge-intensive applications, the work of [12] proposes a Knowledge Model for Hybrid Intelligence applications.
In CommonKADS, a Knowledge Model usually specifies the vocabulary used, i.e. the main classes of the domain (agents, users, paintings in our case) and the processing task(s) performed over such classes (e.g. diagnosis, classification, etc.). These are articulated in three layers: (i) a task layer including the main tasks to be solved, (ii) an inference layer, where these tasks are decomposed into more fine-grained, minimal functionalities (called inferences) and (iii) a domain layer including an ontology with functional terms that serve as inputs and outputs of these functionalities.

Following this structure, the HI Knowledge Model includes a high-level ontology of that describes the most important components in an HI application (actors, processing information, interaction, context), and a taxonomy of HI-relevant tasks (including reasoning, prediction, recognition, actions) that can be achieved by humans and intelligent machines operating in a team. This model is ultimately used as a toolkit to describe and compare HI applications. Similarly, we suggest to use HIKE to measure the HI-ness of the Virtual Guided Tour application. In order to do that, a Knowledge Model of the Virtual Guided Tour needs to be defined, as presented in the next section.

4. The Knowledge Model of the Virtual Guided Tour application

To build the Virtual Guided Tour Knowledge Model, we use a middle-out strategy: we first define the application’s domain ontology (Section 4.1), and then identify the tasks and sub-tasks by linking the domain ontology to the HI one (Section 4.2).

4.1. An Ontology for Virtual Guided Tour

Our Virtual Guided Tour (VGT) ontology represents users (visitors), their actions and interactions with the virtual agent (guide) in a VR environment, following principles of Human-Computer Interaction (HCI) [51] and user modeling [52]. As illustrated in Figure 1, our ontology delineates six key components: Actor, Measures, Environment, Event, Action, and Object, capturing their relationships.

![Figure 1](image.png)

**Figure 1.** Key VGT components: Actor, Measures, Environment, Event, Action, and Object.

The Actor component encapsulates information about users or agents navigating the virtual environment, participating in events, and executing actions. This includes a
comprehensive set of user details such as familiarity with the VR environment, knowledge level about the exhibition, and demographic information (e.g. age, culture, education, gender, and language). Such information allows the agent to adapt to users’ individual preferences and analyze diverse user behaviors to understand underlying patterns. Additionally, details about agents involved in user interactions, such as their capabilities, tasks, and configurations, are also recorded. Each actor performs specific actions during an event, such as the user viewing an area of interest within a painting, or the agent answering users’ questions.

The Action concept can be further categorized into types such as Feedback, Recommendation, Answer, Question, Comment, View, Explanation, Read, and Click. These actions encompass modalities like Gesture, Voice, Gaze, and Facial Expression, each associated with start time, duration, and the actor’s location in the environment during the action. These actions are geared towards achieving a specific Goal [51], driven by the actor’s Intentions, rooted in their overarching goal, giving rise to an Interpretation for an interacting actor.

Inferences about users can be drawn based on various Measures such as gaze-based metrics (fixation count, duration, dwell time, transitions, and scan paths), as well as measures like the number of questions asked and response time. These measures provide valuable insights into information processing, decision-making intricacies, attention patterns, and search strategies [53]. Additionally, they can indicate aspects of user curiosity, attention, and interest [54,55], helping determine emotional states, interests, cognitive characteristics, and personality [56,57,58,59,60,61,62].

The Environment section encompasses various aspects of the virtual environment, including the VR space map, room configurations, domain context, and contextual information about the exhibition. Furthermore, it addresses the goal of the environment, such as conveying the intended message, promoting diversity and inclusion, and ethical considerations related to the virtual environment.

The Object segment provides detailed information about elements within the environment, such as paintings and their textual descriptions. These details include attributes like name, description, story, and coordinates that indicate the object’s position in the environment. Additionally, we cover Area of Interest (AOI) within objects, along with coordinates and associated semantics. These AOIs can represent specific elements (e.g. a person or a building) whose details – such as who he is, what he is proficient in, or the name of the building and its location – can be linked to specific cultural heritage knowledge graphs.

4.2. Inferences and Tasks

In the second phase, we link the VGT ontology to the HI Knowledge Model (refer to Figure 2) to determine the main processes our scenarios deal with.

Firstly, the terms of the domain ontology are mapped to the terms of the ontology of HI knowledge roles. These consist of a few abstract classes and relationships indicating the roles entities play in the reasoning process of any HI application. These terms can be used as inputs and outputs of the minimal processes (inferences) of the application, allowing us to reconstruct the general process in what is called ‘task decomposition.’ By creating these mappings, we achieve two goals: (1) we can describe VGT scenarios in HI terms — including mixed actors, their capabilities, their interactions, and processing
methods, and (2) we can link these terms to the minimal functionalities and derive task decompositions for our VGT scenarios.

Specifically, the classes `vgt:Actor` and `vgt:Capability` are linked respectively to `hi:Actor` and `hi:Capability` through an `rdfs:subClassOf` relationship. Through inheritance, we can describe the agent in terms of the `hi:Interaction` happening between them (through the classes `hi:InteractionTask` and `hi:InteractionMethod`) and the `hi:ProcessingTask` and `hi:ProcessingMethod` they undertake (associated with the class `hi:InformationProcessing`). To map between the classes `vgt:Environment` and `hi:Scenario`, we use the relationship `skos:broader` given its broader semantics. This allows us to describe `hi:Domain`, `hi:Context`, `hi:EndGoal`, and potential `hi:EthicalConsiderations`, if needed. Table 1 summarizes the knowledge roles and tasks in our scenarios. For example, in Scenario 1 (S1), the processing method is “Multimodal Analysis & Neurosymbolic Reasoning,” and the identified processing task is “Reasoning.” This reflects the cognitive process the virtual agent uses to interpret Sarah’s multimodal cues and customize the interaction accordingly.

Table 1. Knowledge roles of the Virtual Guided Tour. The classes `hi:Actor` and `hi:Domain` are omitted, as all scenarios deal with Sarah and the virtual agent, operating in a Virtual Guided Tour application.

<table>
<thead>
<tr>
<th>hi:EndGoal</th>
<th>hi:ProcessingMethod</th>
<th>hi:ProcessingTask</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1  Multimodal Interaction</td>
<td>Multimodal Analysis &amp; Neurosymbolic</td>
<td>Reasoning</td>
</tr>
<tr>
<td>S2  Real-time State Understanding</td>
<td>Theory of Mind</td>
<td>Generate</td>
</tr>
<tr>
<td>S3  Memory Integration</td>
<td>Symbolic techniques</td>
<td>Transform</td>
</tr>
<tr>
<td>S4  Perception Building</td>
<td>Multimodal Analysis &amp; Theory of Mind</td>
<td>Learning</td>
</tr>
<tr>
<td>S5  Explainability</td>
<td>Statistical techniques</td>
<td>Generate</td>
</tr>
<tr>
<td>S6  Collaborative Pursuit of Goals</td>
<td>Reinforcement Learning</td>
<td>Learning, Generate</td>
</tr>
</tbody>
</table>

The next step is to determine a set of relevant fine- and coarse-grained processes that can help us build the task layer of the Virtual Guided Tour application. Recall that, in a task decomposition process, coarse-grained tasks are decomposed into smaller, fine-grained sub-tasks hierarchically, and the leaves at the lowest level are the primitive processes (inferences) linked to the terms of the ontology (the inputs and outputs of such processes). As suggested by HIKE, our primitives include (i) “transfer functions” that allow communication between agents (`obtain, receive, present, provide` a given input/output), (ii)
classical KE processes (abstract, select, transform certain data), and (iii) VR-specific processes such as perceive through a VR headset or interact with virtual objects. We also maintain the distinction between static inputs (e.g. the cultural heritage Knowledge Graph, a room in the VR exhibition) and dynamic inputs (e.g. an explanation to Sarah, Sarah’s eye gaze, etc.). Finally, our scenarios rely on both the physical and the computational component, hence we use all the high-level tasks proposed by the HIKE toolkit, e.g. recognition of Sarah’s mood, prediction of the next painting to recommend, reasoning over Sarah’s interests (deductively based on historical data, or inductively based on her current visual attention), and action (including both cognitive actions based on Theory of Mind or memory and physical actions such as gaze or movements).

5. Validation

The VGT Knowledge Model is now used to answer our research questions: namely, whether and which HI aspects can be identified using the HIKE toolkit (Section 5.1), and how to use HIKE to evaluate HI-ness of our scenarios (Section 5.2).

5.1. RQ1 : Task Decomposition of VGT scenarios

Our goal here is to assess the effectiveness of HIKE in identifying aspects of Hybrid Intelligence within the VGT scenarios. To achieve this, we use the VGT Knowledge Model to represent our scenarios in the form of processing workflows\(^3\). We then compare the number of classical KE tasks vs. the HI tasks identified by the HIKE toolkit within these workflows. Given space constraints, presenting decompositions for all six tasks is impractical. Therefore, for illustrative purposes, Figure 3 presents S1, S2, S3, and readers are encouraged to refer to our online repository\(^4\) to access the complete set of the material. Additionally, the high-level tasks of the scenarios are outlined in Table 2.

We recognize in our scenarios recurring tasks such as Recognition, Communication, Monitoring, Modeling and Adaptation. Additionally, we observe Recommendation, Recollection, Explanation, Decision Making, and Perception in all scenarios but S2. In Table 2, we mark HI-specific tasks as ✓✓ (Communication, Explanation, Adaptation, Perception, Collaborative Decision Making), open-ended knowledge engineering tasks as ✓ (Recollection, Recognition, Recommendation), and classical CommonKADS tasks without checkmark (Monitoring, Modeling). All scenarios include at least one HI task, which not only demonstrates the existence of HI-specific tasks, but also reflects the HI concepts of adaptability, collaborativeness, and explainability at the core of the HI agenda\(^1\).

Additionally, the decompositions in Figure 3 reveal new, recurrent elements that are typical of hybrid environment: sub-tasks such as perceive multimodal inputs through a VR headset (all scenarios), adapt behavior (S2, S3), and update memory (S1, S6); inputs/outputs types such as multimodal (VR) actions (all scenarios) and Theory of Mind

---

\(^3\)Note the recommended UML notation: ovals (inferences), rounded rectangles (transfer functions), full squares (dynamic inputs/outputs), horizontal lines (static inputs/outputs), dotted arrows (lists), dashed boxes (more general tasks).

(a) S1: Multimodal Interaction.

(b) S2: Real-time Understanding of User States.  (c) S3: Memory Integration for Personalization.

Figure 3. Task Decomposition for Scenarios S1, S2 and S3.

(S2, S6). These new aspects also confirm the suitability of the task decomposition to identify HI-ness of an application, but suggest that additional terms might be needed within HIKE. For example, we face challenges in explicitly attributing tasks to specific actors (agent, user, or the creator of the virtual exhibition), hampering the use of inferences such as alter virtual environment or interact with virtual object.

5.2. RQ2: Assessing HI-ness of VGT Scenarios

To address our second question, we focus on assessing the level of HI-ness in terms of synergy between humans and machines (i.e. the stronger the HI, the greater the synergy in collaboration). In Table 2, we describe the single tasks we identified in each scenario, and mark them with a Weak HI level if a task involves a mixed team without collaboration toward the same goal, and a Strong HI level if the teams actually create synergy. Ultimately, we provide an overall percentage to assess the HI-ness in the entire scenario.

We observe that the Modeling tasks (S1, S3) primarily involve data storage and retrieval of a single agent, representing minimal synergy and hence a weak level of Hybrid Intelligence. Similarly, tasks categorized under Recognition (S1, S2, S5, S6) primarily focus on pattern recognition or acquiring user information, indicating slightly increased collaboration within the team, but still demonstrating weak HI due to a primarily data-driven interaction rather than true collaboration.
Table 2. VGT scenarios, task description, task HI-ness specificity and level, and total scenario HI-ness

<table>
<thead>
<tr>
<th>S*</th>
<th>Task</th>
<th>Task Description</th>
<th>HI Task</th>
<th>HI Level</th>
<th>HI-ness</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Modeling</td>
<td>Storing user VR multimodal actions, user interest, agent actions, and communication style of the agent</td>
<td></td>
<td>Weak</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>Recognition</td>
<td>Determining user interest based on VR action and updating the user model</td>
<td>✓</td>
<td>Weak</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Recommendation</td>
<td>Selecting agent actions based on agent and user models</td>
<td>✓</td>
<td>Strong</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Communication</td>
<td>Providing knowledge to the user</td>
<td>✓ ✓</td>
<td>Weak</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monitoring</td>
<td>Evaluating user interest based on current user actions after providing knowledge</td>
<td></td>
<td>Weak</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td>Recognition</td>
<td>Determining the user’s cognitive state through Theory of Mind techniques</td>
<td>✓</td>
<td>Weak</td>
<td>33%</td>
</tr>
<tr>
<td></td>
<td>Adaptation</td>
<td>Adapting agent actions based on recognized user state</td>
<td>✓ ✓</td>
<td>Strong</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monitoring</td>
<td>Evaluating user engagement based on current actions after providing knowledge</td>
<td></td>
<td>Weak</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>Recollection</td>
<td>Recollecting information about the user, such as interests, previous interactions</td>
<td>✓</td>
<td>Strong</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>Adaptation</td>
<td>Adapting agent utterances based on collective patterns of users stored in memory</td>
<td>✓ ✓</td>
<td>Strong</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Communication</td>
<td>Providing knowledge to the user</td>
<td>✓ ✓</td>
<td>Weak</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Modeling</td>
<td>Storing agent actions</td>
<td></td>
<td>Weak</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>Perception</td>
<td>Shaping perception about the agent’s capabilities</td>
<td>✓ ✓</td>
<td>Strong</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Communication</td>
<td>Providing information about capability to the user</td>
<td>✓ ✓</td>
<td>Strong</td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>Recognition</td>
<td>Recognizing the intent of the user’s question</td>
<td>✓</td>
<td>Weak</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>Explanation</td>
<td>Providing the reason behind the agent’s action</td>
<td>✓ ✓</td>
<td>Strong</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>Recognition</td>
<td>Recognizing the goal of the user based on their actions</td>
<td>✓</td>
<td>Weak</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>Decision Making</td>
<td>Deciding the learning path based on system and user objectives</td>
<td>✓</td>
<td>Strong</td>
<td></td>
</tr>
</tbody>
</table>

To enhance these weaker tasks and bolster HI, it is essential to incorporate mechanisms where the agent seeks clarification or confirmation from the user when uncertain about its decisions regarding user interests or mental state. By integrating user confirmation, ideally coupled with clear explanations, the agent can refine its decision-making strategies, thereby fostering better judgment in analogous situations in the future. This interactive process ultimately culminates in a heightened level of hybrid intelligence, characterized by enhanced capabilities in terms of collaboration, adaptivity, and explainability.

In contrast, the Recommendation tasks (S1), where the agent chooses actions based on both user and agent models, stand out as a strong HI component. Here, the agent refines its actions and adapts information to the user, showcasing a high degree of synergy. The Adaptation tasks (S2, S3) also show strong HI-ness, by dynamically adapting agent actions based on user states and collective patterns from memory. Conversely, the Com-
Communication (S1, S3) and Monitoring tasks (S1, S2) lean towards weak HI-ness, emphasizing one-way information transfer and post-information evaluation. Communication in S4 can be considered strong as the agent provides information about its capabilities and limitations (e.g., “I can only answer factual questions”, “I am an expert in 17th-century paintings” or “I am not familiar with the fashion style back then”). This helps the user understand the agent’s capabilities and foster Team Awareness – an important HI capability. Examining the Explanation task (S5) reveals a strong HI aspect, as providing explanations for the agent’s actions involves building trust and understanding. Similarly, the Decision Making task (S6) demonstrates strong HI-ness by considering both agent and user goals, fostering collaboration and mutual satisfaction.

Overall, 4 out of 6 scenarios present at least half of their tasks with strong HI-ness (50% or above). The weakest scenarios (S1, S2) involve classical KE tasks, but still include HI aspects such as adaptivity, Theory of Mind reasoning and user-agent shared understanding. In the scenarios with 50% HI-ness, collaboration through communication appear to be the key HI component. S4, involving building perception jointly through communication, finally appear to be the scenario with the strongest HI-ness.

Notice that HIKE serves not only as an analytical tool, but also as an active method to recommend adaptations and improvements of one’s scenarios. For instance, S1 includes a HI task (Communication) whose the implementation is still evaluated as weak. This could suggest that the design of the Communication module should be modified, not only to increase the VGT’s HI-ness, but potentially the overall quality of an application. Similarly, HIKE could help improving scenarios involving non-specific HI tasks. In the case of e.g. Recognition in S6, the curator might decide strengthen the role of the user by, for instance, using different methods for recognition (MTurks, explicit user-based models, etc.). This analysis confirms the idea that a Knowledge Engineering method such as HIKE can be used to assess, and possibly improve, the HI-ness of an application.

6. Conclusions

In this paper, we employed Knowledge Engineering to assess the level of HI-ness of a Virtual Guided Tour application designed for the Virtual Heritage domain. We described 6 different scenarios in the form of workflows of tasks and input/outputs, and identified the specific Hybrid Intelligence tasks performed within each scenario. We ultimately suggest to assess the HI-ness of the scenarios by (i) comparing the number of classical Knowledge Engineering tasks vs. HI-specific tasks, and (ii) measuring the strength of synergy achieved by the mixed team in the various tasks. Our study reveals interesting insights, including the usability of the Hybrid Intelligence Knowledge Engineering (HIKE) toolkit to identify HI aspects in an existing application, the possibility of measuring and improving its level of HI-ness, as well as the need of improving certain modeling aspects within HIKE. Furthermore, our findings indicate that HIKE can function as a design tool to facilitate the creation of HI scenarios. Future work will revolve around improving HIKE through validation with new scenarios, strengthening the concept of HI-ness, and standardization of the HI task templates (intended as recurrent combination of tasks and input/outputs) across HI initiatives. Additionally, the paper mainly focuses on the AI agent, with Table 2 presenting tasks from the agent’s perspective. For future work, our aim is to investigate how Table 2 would appear when structured from the human task perspective.
References


Liu D. Knowledge Graph Driven Conversational Virtual Museum Guide. University of Twente; 2021.


Evaluating Generative AI Incidents: An Exploratory Vignette Study on the Role of Trust, Attitude and AI Literacy

Esther S. KOX\textsuperscript{a,b,1} and Beatrice BERETTA\textsuperscript{a}
\textsuperscript{a}University of Twente, Enschede, The Netherlands
\textsuperscript{b}TNO, Soesterberg, The Netherlands
ORCiD ID: Esther Kox https://orcid.org/0000-0002-2512-5665

Abstract. Generative AI presents vast opportunities but also risks. Misuse, whether intentional or not, can lead to significant “real-world” consequences. We presented subjects (n=139) with five vignettes describing incidents involving generative AI. We explored the relationship between their level of AI literacy, attitude towards AI, trust in AI chatbots, and people’s reactions to the vignettes. Attitude and trust, measured before and after the vignettes, declined significantly. However, these changes as well as the reactions to the vignettes were unrelated to AI literacy. Yet, higher AI literacy was associated with more frequent use of AI chatbots, higher trust and more positive attitudes towards AI. So while AI literacy appeared to be related to the general perceptions and usage of generative AI, it was not linked to the evaluation of incidents involving generative AI. The implications for trust calibration and appropriate reliance are discussed.

Keywords. LLMs; AI chatbots; trust; attitude; AI literacy

1. Introduction

Artificial Intelligence (AI) technology has revolutionized the way we handle information. Especially generative AI models (e.g., ChatGPT) are seen as one of the most disruptive technological breakthroughs in recent years [1,2]. Since its release end of 2022, ChatGPT, an AI chatbot developed by OpenAI, has gradually gained popularity and has changed how people perceive and interact with AI [3,4]. Models like ChatGPT are Large Language Models (LLMs); a type of AI designed to understand and generate language in a way that mimics human language processing abilities [5]. These models are trained using immense amounts of publicly available text from the internet, exposing the model to a wide range of topics, writing styles, and linguistic patterns [5], which enables them to capture the nuances of human language and to produce highly coherent and human-like responses [2,6,7]. These qualities make them perfect for conversational use and have contributed greatly to their popularity. Generative AI offers many potential and realized benefits for people, organizations and society in a wide range of sectors [2,8].

\textsuperscript{1}Esther S. Kox, esther.kox@tno.nl
However, the use of generative AI also carries several risks [2,6]. First, generative AI can produce harmful and inappropriate content (e.g., discriminative content, promotion of harmful ideologies) [2]. Both the model's training data and its lack of contextual understanding can lead it to generate content that is considered inappropriate or culturally insensitive. Second, generative AI can fabricate fictitious or erroneous content with a high level of plausibility [9]. In literature, this phenomenon is often referred to as “hallucination” or “confabulation” [10,11], but we will use the non-anthropomorphic term fabrication [2,12]. Third, the increasing difficulty in determining the authenticity of information and media (e.g., Deepfake) [2]. Fourth, training data for algorithms often contain biases, unfairly favouring or disadvantaging certain individuals or groups, which can seep into the model's output, sustaining societal biases [1,2].

These issues can cause a wide range of real-life consequences [6,13]. For example, biased algorithms can, when deployed and acted upon, exacerbate existing societal inequality. Also, if patients, for example, rely on fabricated information for medical advice, it could result in life-threatening situations [5,6,14]. Alternatively, the manipulation of audio-visual media can exacerbate the far-reaching effects of misinformation on social media, such as polarisation [15]. The possible instances of intentional and unintentional misuse are numerous [5]. To mitigate the risks and maximize the benefits associated with generative AI, people must understand how AI works and is applied, so that they can trust in and rely on it appropriately.

1.1. Trust

During interaction, people continuously adjust their level of trust in an artificial agent based on their ongoing interactions and experiences with the aim to align the perceived trustworthiness of an agent with their actual trustworthiness (i.e., trust calibration) [16]. Consistent with that objective, people are increasingly (made) aware of the fact that AI chatbots can provide misleading information that could harm user’s interests or well-being (e.g., false medical, legal, or financial advice) [6]. For instance, OpenAI added a warning to ChatGPT’s main screen: “ChatGPT can make mistakes. Verify important information.” Nevertheless people seem to trust in and rely on AI more than they should.

Prior work shows how people are often misled by incorrect AI predictions and how they would, in some cases, make better decisions on their own [17,18]. Trusting false AI generated output can have major societal implications. For instance, a lawyer depended on ChatGPT to draft a motion replete with fabricated case law, because he “did not comprehend that ChatGPT could fabricate cases” [19] (see vignette ‘Fabrication’, Table 1). People often tend to follow the advice of automation, without verifying it, because they consider a machine as infallible (i.e., automation bias) [7,20]. Adding a warning that ChatGPT can make mistakes is an attempt to prevent overreliance. However, the effect of the warning is likely to be nullified by the convincing and seemingly sophisticated output that ChatGPT generates.

ChatGPT’s ability to generate highly coherent answers “can fool us into thinking that they understand more than they do” [21]. ChatGPT’s output seems highly plausible and intelligent, but lacks comprehension [22]. However, the human-like way of communicating triggers people’s tendency to attribute humanlike capabilities to non-human entities (i.e., anthropomorphism) [23]. People tend to base their level of trust on attributed characteristics rather than on actual experiences with the agent itself [24],
creating a discrepancy between the perception and its actual capabilities [6]. As such, anthropomorphism can lead to misplaced trust and inappropriate reliance.

Some scholars therefore argue that anthropomorphic features in the design of artificial agents should be avoided [17, 24, 25]. However, despite the potential risks, the human-like responses are exactly what contributed to the success and ease-of-use of AI chatbots. But not all people are equally susceptible to anthropomorphic cues [23, 26]. Researchers have proposed that people who lack AI literacy tend to anthropomorphize AI-agents more [6]. People with a limited understanding of AI may more easily fall for the illusion of intelligence and overestimate it based on superficial interactions. Increasing people’s AI literacy could mitigate the risks associated with anthropomorphic design, while holding on to its benefits.

1.2. AI literacy

AI literacy is defined as a broad set of skills that enable individuals to recognize everyday applications of AI, know the basic functions of AI and understand how to use AI effectively in daily life [27]. For many, AI is still a “black box” with difficult to determine opportunities and risks [28]. AI literacy enables individuals to make informed decisions about AI [29] and help people to gauge when it is appropriate to rely on it [8]. For example, research shows that clinicians with higher AI literacy were less likely to rely on incorrect medical AI recommendations than clinicians with lower AI literacy [18]. AI literacy is thought to counter biases that are known to interfere with accurate trust calibration, appropriate reliance and effective decision-making (e.g., anthropomorphism, automation bias). As such, understanding AI’s strengths and weaknesses is deemed crucial for mitigating instances of misuse and deception like mentioned earlier [6, 16, 30]. More research is needed to determine how AI literacy affects people’s trust in, attitude towards and perceptions of AI across different contexts.

1.3. Current study

The aim of this study was to explore how people respond to scenario’s describing incidents involving generative AI and whether these responses are influenced by an individual’s level of AI literacy. It was expected that trust in and attitudes towards AI of people with higher level of AI literacy would be less affected by the vignettes in comparison to people with lower AI literacy. This stems from the idea that higher AI literacy results in more insight into AI associated risks and the consequences described in the scenarios and thus a more grounded and robust level of trust.

2. Method

2.1. Participants and design

In this non-experimental exploratory questionnaire study, we presented 139 subjects, ages ranged from 18 to 65 years old (M = 32.74, SD = 13.17), with five vignettes describing incidents involving generative AI (Table 1). We examined potential correlations between an individual’s AI literacy, attitude towards AI, trust in AI
chatbots, and their reactions to the vignettes (see 2.4). Attitude and trust were measured before and after the vignettes. Ethical approval was attained from the Ethics Committee of University of Twente’s Behavioural, Management and Social Sciences’ (BMS) faculty. A diverse group was gathered through voluntary sampling techniques via online platforms, such as Survey Circle and the university’s SONA system. A total of 185 individuals initially participated, but our final dataset comprised 139 participants due to incomplete responses and exclusions. Specifically, 33 were incomplete, four were excluded for a completion time under 5 minutes, and nine were excluded for unfamiliarity with AI chatbots.

### Table 1. The five vignettes presented in the study.

<table>
<thead>
<tr>
<th>Type</th>
<th>Vignette</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harmful content</td>
<td>An organization that supports people with eating disorders introduced an AI chatbot as a tool that could offer prevention strategies for people with eating disorders, such as anorexia and bulimia. However recently, users started sharing screenshots of their experience with the chatbot via social media. They reported that the bot provided harmful advice, recommending weight loss, counting calories, and measuring body fat; behaviours that could potentially exacerbate eating disorders. Patients, families, doctors and other experts on eating disorders were left stunned and bewildered about how a chatbot designed to help people with eating disorders could end up dispensing diet tips instead.</td>
<td>[31]</td>
</tr>
<tr>
<td>Inappropriate content</td>
<td>A public transport company wanted to create a funny commercial. It decides to commission an advertisement from an AI marketing system that uses a play on the word riding. The resulting ad, Figure 1, causes shock and outrage among members of the public.</td>
<td>[32]</td>
</tr>
<tr>
<td>Fabrication of sources</td>
<td>A lawyer at a respected firm used an AI chatbot to find historic cases relevant to his client’s lawsuit. The chatbot came up with a list of twelve cases. Later in court it turns out that the chatbots findings were completely made up. Court documents show that half of the submitted cases appear to be bogus judicial decisions with bogus quotes and bogus internal citations.</td>
<td>[19,33]</td>
</tr>
<tr>
<td>Plagiarism</td>
<td>A record label hired an AI songwriter to write lyrics for famous musicians. The AI songwriter has written lyrics for dozens of songs in the past year. However, a journalist later discovers that the AI songwriter has been plagiarizing lyrics from lesser-known artists. Many artists are outraged when they learn about the news.</td>
<td>[32]</td>
</tr>
<tr>
<td>Bias</td>
<td>To improve their admission process, a university began using a new AI machine-learning system to help make decisions about who gets into its Ph.D. program - and who doesn’t. The algorithm evaluates grades, test scores, and recommendation letters of applicants. An audit revealed that the new algorithm is biased against minority applicants. Critics concerned about diversity, equity and fairness in admissions are angry and say the system exacerbates existing inequality in the field.</td>
<td>[32,34]</td>
</tr>
</tbody>
</table>

### 2.2 Task and procedure

Data was gathered via the online survey platform Qualtrics. Participants were first presented with information about the study and an informed consent form. Upon agreeing to participate, demographics, experience with AI chatbots, AI literacy, Trust and Attitude were administered (see 2.4). After that, participants were presented with five vignettes. The vignettes were introduced with the text: “In the next section you will be presented with a series of scenario’s about different applications of and actions by AI. Please read the scenario’s carefully and answer the questions.”. After each vignette, their perceptions about the actions described in the vignette were assessed. The order of the five vignettes was randomized between participants. After the vignettes, Trust in AI Chatbots and Attitude towards AI were measured again. Finally,
we inquired about their intentions to continue using AI chatbots. Two vignettes (i.e., Inappropriate content and Plagiarism) were hypothetical and adopted from [32]. Two vignettes (i.e., Harmful content and Fabrication of Sources) were based on recent news articles. The final vignette (i.e. Bias) was a combination of both.

![Figure 1. Picture shown with the Inappropriate content vignette (from [32])]  

2.3 Measures

**Experience with AI chatbots.** Participants were asked if they knew AI chatbots (e.g. ChatGPT) (yes; no) and if yes, if they have used it and if so, how often they use it (never, once a year, once a month, almost every day) [35].

**AI literacy** was measured the MAILS - Meta AI Literacy Scale [36] (Cronbach’s α = 0.94), consisting of four subscales: Use & Apply AI (“I can use AI applications to make my everyday life easier.”) (α = 0.96); Know & understand AI (e.g. “I can assess what the limitations and opportunities of using AI are.”) (α = 0.96); Detect AI (e.g., “I can tell if I am dealing with an application based on AI.”) (α = 0.86); and AI Ethics (e.g., “I can incorporate ethical considerations when deciding whether to use data provided by an AI.”) (α = 0.88). Participants rated their own abilities on 18-items on a scale from 0 (i.e., hardly or not at all pronounced) to 10 (i.e., very well or almost perfectly pronounced). All subscales demonstrated strong internal consistency, as indicated by Cronbach’s alpha coefficients.

**Trust in AI Chatbots** was measured using the 12-item human-computer trust scale [37] (e.g., “I think that AI Chatbots perform their role as personal assistant very well”) (prior: α = 0.78, post: α = 0.95). Participants rated their accordance with the statements on a 5-point Likert scale ranging from “Strongly disagree” to “Strongly agree”.

**Attitude towards AI** was measured with four items [32]. Participants were first asked how they weighed the potential risks and benefits of AI and then to rate their accordance with three statements “AI makes me feel… worried/ hopeful/ angry” on a 6-point scale (i.e., No; Yes, just a little; Yes, slightly; Yes, moderately; Yes, quite; Yes, very). In our analysis, the variable ‘Attitude towards AI’ represents the calculated mean of the scores on the latter three items, where worried and angry are reversed so that a higher score represents a more positive attitude. The ordinal item about the weighed risks and benefits was analysed separately.

**Perceptions** consisted of four items. Participants rated how surprising / harmful / morally wrong / emotionally distressing they found the described actions in the vignette, on a 5-point Likert scale ranging from “None at all” to “A great deal”. In our analysis, the aggregated variable ‘Perception’ represents the calculated mean of how harmful, morally wrong and emotionally distressing they found the described action. The surprising item was not included in the aggregated measure since it does not gauge an affective response. It was included to assess participants’ familiarity with the examples.
Lastly, **continuance intention** was measured with two items: “I plan to keep using AI chatbots” and “I want to continue using AI chatbots” on a 5-point Likert scale ranging from “Strongly disagree” to “Strongly agree”. (not applicable option was available). [38].

### 3. Results

#### 3.1. Descriptives

Descriptives per variable are shown in Table 2. The matrix also shows that (1) AI literacy (var 7 to 11) is not correlated with the perceptions of the vignettes (var 12 to 16), (2) AI literacy (total, var 11), frequency of use, trust in AI chatbots, attitude towards AI and continuance intentions are all positively correlated, (3) age is not correlated with AI literacy.

|        | M      | SD    | 1    | 2    | 3    | 4    | 5    | 6    | 7    | 8    | 9    | 10   | H    | 12   | 13   | 14   | 15   | 16   | 17   |
|--------|--------|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Age    | 32.7   | 13.2  | .19  | .01  | .06  | .10  | .03  | .09  | .20  | .11  | .14  | .09  | .03  | .00  | .03  | .29  | .20  | .01  |
| Freq. of use | 3.42 | 1.14 | 1    | .37  | .32  | .22  | .21  | .60  | .18  | .15  | .01  | .37  | .37  | .06  | .09  | .17  | .05  | .02  | .55  |
| Trust (prior) | 2.76 | 0.56 | 1    | .81  | .50  | .45  | .40  | .17  | .22  | .04  | .29  | .13  | .11  | .18  | .21  | .13  | .38  |
| Trust (post) | 2.57 | 0.64 | 1    | .50  | .51  | .43  | .16  | .23  | .04  | .30  | .08  | .05  | .14  | .27  | .18  | .44  |
| Attitude (prior) | 3.71 | 0.69 | 1    | .79  | .35  | .17  | .18  | .08  | .25  | .11  | .05  | .06  | .12  | .08  | .40  |
| Attitude (post) | 3.49 | 0.74 | 1    | .32  | .16  | .22  | .06  | .24  | .17  | .18  | .15  | .23  | .18  | .49  |
| AI lit. (apply) | 5.45 | 2.45 | 1    | .46  | .42  | .26  | .78  | .10  | .14  | .11  | .07  | .05  | .11  | .14  | .18  |
| AI lit. (know) | 6.09 | 2.10 | 1    | .61  | .69  | .87  | .11  | .07  | .05  | .11  | .14  | .18  |
| AI lit. (detect) | 5.49 | 2.21 | 1    | .56  | .75  | .03  | .02  | .02  | .01  | .05  | .16  |
| AI lit. (ethics) | 5.82 | 2.17 | 1    | .71  | .02  | .04  | .09  | .04  | .03  | .01  |
| AI lit. (total) | 5.73 | 1.78 | 1    | .08  | .08  | .06  | .07  | .09  | .31  |
| 12 Resp. Plagiarism | 3.53 | 0.98 | 1    | .37  | .57  | .53  | .67  | .05  |
| 13 Resp. Inappropriate | 3.27 | 1.20 | 1    | .38  | .39  | .34  | .05  |
| 14 Resp. Fabrication | 3.74 | 0.91 | 1    | .60  | .52  | .04  |
| 15 Resp. Harmful | 3.94 | 0.95 | 1    | .67  | .12  |
| 16 Resp. Bias | 3.96 | 0.91 | 1    | .01  |
| 17 Cont. intention | 3.97 | 0.94 | 1    |

*Correlation is significant at the 0.05 level (2-tailed).
**Correlation is significant at the 0.01 level (2-tailed).

### 3.2. Trust & Attitude before and after

A paired-samples t-test revealed a significant difference between Trust in AI Chatbots before (M1 = 2.76, SD1 = 0.56) and after (M2 = 2.57, SD2 = 0.64) the vignettes, t(138) = 6.60, p < .001. A Cohen’s d of 0.38 suggests a medium-sized effect.

A paired-samples t-test revealed a significant difference between Attitude towards AI before (M1 = 3.71, SD1 = 0.69) and after (M2 = 3.49, SD2 = 0.74) exposure to the vignettes, t(138) = 5.70, p < .001. A Cohen’s d of 0.47 suggests a medium-sized effect.
To explore whether differences in trust and attitude prior and post exposure to the vignettes were related to an individual’s AI literacy, the respective differences (deltas) between the two measurement points (i.e., pre and post vignettes) of Trust and Attitude (e.g., Trust_{post} – Trust_{pre}) were calculated as measures of observed change. Then, we calculated the correlations between AI literacy (total) and the two delta values (i.e., ΔTrust, ΔAttitude). The correlation between ΔTrust and AI literacy was non-significant, r(139) = 0.09, *p* = .319. The correlation between Δ Attitude and AI literacy was also non-significant, r(139) = 0.01, *p* = .877. This suggests no relation between AI literacy and the change in Trust and Attitude before and after exposure to the vignettes.

### 3.3. Perceptions & AI literacy

Finally, Pearson’s correlation coefficients were calculated between the subscales of the AI literacy questionnaire and the separate perception items per vignette (i.e., Surprising, Harmful, Morally wrong, Emotionally distressing). Most correlations were non-significant. However, four perceptions were weakly negatively correlated with the “Know & Understand” subscale of AI literacy, namely Morally Wrong in Plagiarism (r(139) = -.17, *p* = .041), Surprising (r(139) = -.21, *p* = .013) and Harmful (r = -.17, *p* = .045) in Inappropriate content and Morally Wrong in Bias (r(139) = -.17, *p* = .040). One perception (i.e., Surprise in Harmful content) was negatively correlated with the “AI Ethics” subscale (r(139) = -.19, *p* = .026).

### 4. Discussion

In line with prior findings, our results showed that higher AI literacy was associated with higher trust and a more positive attitude towards AI [8]. These factors were further linked to higher frequency of use and increased intentions for continued use. Given the correlational nature of our study, we remain uncertain about cause and effect, as well as possible bidirectional causal relationships. People with higher trust might use AI more, but frequent use might also foster trust. Also, people who see the benefits of AI are more likely to use it and try to comprehend it, thereby increasing their AI literacy. Also, Yet, learning more about AI can also foster a more positive attitude towards it. The positive correlation between AI literacy and attitude and trust can also be seen as somewhat surprising, as some have proposed AI literacy as the antidote for overreliance [18,39]. However, people with higher AI literacy might have a positive and trustful perception of AI; yet largely based on knowledge and experience, rather than grounded on gut feeling and biases. As such, promoting AI literacy is still seen as a means to ensure calibrated trust. Further research is needed to provide more clarity on which variable influences the other and the potential effect of third variables.

Further, AI literacy was unrelated to ones affective responses to incidents with AI. First, AI literacy was unrelated to how morally wrong, harmful, or emotionally distressing people found the incidents described in the vignettes. Second, the reductions in trust and attitude were also unrelated to AI literacy. How people evaluated the incidents may be more closely linked to how people perceive (moral) incidents in general, also those not involving AI. For instance, the extent to which people might find copyright infringement or discriminatory biased decision-making (as described in some of the vignettes) morally wrong, harmful or emotionally distressing has perhaps more to do with ones general norms and values than with their level of AI literacy.
We did observe significant relations between the AI literacy subscales and perceptions separately. Notably, “Know & Understand AI” was negatively correlated with some of the perceptions of the described incidents, suggesting that, in some cases, understanding the limitations and opportunities of using AI was associated with a somewhat milder response to the incidents. However, for the most part, understanding that and how such incidents can occur on a cognitive level did not inherently lead to greater forgiveness for the resulting damage.

While unrelated to AI literacy, the reductions in trust and attitude after the vignettes do indicate that showing the possible disadvantageous outcomes of using generative AI has a significant effect on how people perceive AI. People generally adapt and aim to calibrate their level of trust as they learn more about artificial agents. As this fairly new technology emerges and spreads in society, people will continuously learn about its capabilities and limitations across different context and adjust their trust and reliance accordingly. Fostering a calibrated level of trust is crucial to minimize the risks and to maximize the benefits of new technology [16,40,41].

4.1. Limitations & future research

One limitation of this study is that the perception measure (i.e., how do you perceive the described actions in the vignette?) did not differentiate between the main actor and the AI system. Participants may have answered with the main actor in mind rather than the AI system, what would explain the lack of coherence between these perceptions and AI literacy. Second, we used a self-reported AI literacy that covered a wide a range of skills [36,42]. However, its main focus was on cognitive and ethical aspects and less on people’s attitudinal characteristics to learn AI [43]. Attitudinal aspects might better explain people’s affective responses to AI incidents. Future studies including attitudinal aspects of AI literacy as well as potential differences between AI literacy performance and self-reported AI literacy would be worthwhile. Also, in future research we would include perceived anthropomorphism. Researchers have suggested a link between AI literacy with the tendency to anthropomorphize artificial agents [6]. Further investigations are needed to study the link between AI literacy and the perceived anthropomorphism of AI chatbots.

4.2. Conclusion

Instances of intentional and unintentional misuse of generative AI indicate that many people lack a thorough understanding of its limitations in particular [19]. Half of our vignettes, describing incidents involving generative AI with societal implications, were based on recent news articles, highlighting the urgency of this issue. AI literacy is seen as a prerequisite for people’s ability to determine when it is appropriate to trust in and rely on generative AI [28], which can help minimize the risks of this promising new technology. The present study has gone some way towards enhancing our understanding of how AI literacy is related to people’s trust in, attitude towards and perceptions of AI across different contexts. While AI literacy appeared to be related to the general perceptions and usage of generative AI, it was not linked to the evaluation of incidents involving generative AI. Yet, we found that participants’ attitudes towards AI and trust in AI chatbots declined significantly after reading the vignettes. This suggests that informing people about the possible disadvantageous outcomes of using generative AI can change how people perceive and trust AI.
References


sanctions.html


[35] Montag C, Ali R. Can we assess attitudes toward AI with single items? Associations with existing attitudes toward AI measures and trust in ChatGPT in two German speaking samples. 2023;


Unreflected Acceptance - Investigating the Negative Consequences of ChatGPT-Assisted Problem Solving in Physics Education

Lars KRUPP, Steffen STEINERT, Maximilian KIEFER-EMMANOUILIDIS, Karina E. AVILA, Paul LUKOWICZ, Jochen KUHN, Stefan KÜCHEMANN, and Jakob KARULUS

German Research Center for Artificial Intelligence (DFKI)
RPTU Kaiserslautern-Landau
LMU Munich

Abstract. The general availability of large language models and thus unrestricted usage in sensitive areas of everyday life, such as education, remains a major debate. We argue that employing generative artificial intelligence (AI) tools warrants informed usage and examined their impact on problem solving strategies in higher education. In a study, students with a background in physics were assigned to solve physics exercises, with one group having access to an internet search engine (N=12) and the other group being allowed unrestricted use of ChatGPT (N=27). We evaluated their performance, strategies, and interaction with the provided tools. Our results showed that nearly half of the solutions provided with the support of ChatGPT were mistakenly assumed to be correct by students, indicating that they overly trusted ChatGPT even in their field of expertise. Likewise, in 42% of cases, students used copy & paste to query ChatGPT — an approach only used in 4% of search engine queries — highlighting the stark differences in interaction behavior between the groups and indicating limited task reflection when using ChatGPT. In our work, we demonstrated a need to (1) guide students on how to interact with LLMs and (2) create awareness of potential shortcomings for users.

Keywords. ChatGPT, Large Language Models, Education, Physics

1. Introduction

LLMs have been omnipresent in media and the public eye since November 2022 when ChatGPT was first presented [1]. With one of the fastest growing user bases ever measured for any application [2,3], it is difficult to estimate its future impact on every aspect of our daily lives.

Especially in sensitive areas, such as education, easily accessible information — true or false — poses challenges for educators and students alike. Recent discussions
around ChatGPT often involve its use as an AI support tool in assignments, for homework and in the classroom. Despite research advances, it is still unclear how LLMs, such as ChatGPT, can meaningfully support students in educational contexts [4,5]. To properly design methods that allow for informed usage of these systems, we need to investigate how the users — in our case students — interact with those AI tools and how their usage influences the students’ decision making.

LLMs are predictive models that predict the most probable next token based on a series of previously seen tokens they have already seen. As a result, they excel at tasks such as brainstorming [6], writing [7], translation [8], and even programming [9]. Contrarily, disciplines that rely heavily on calculations and reasoning prove more challenging for LLMs [10]. Potentially leading to unforeseen or even negative consequences for students, like incorrect homework [4] or learning an incorrect explanation of a concept. Yet, it remains unclear how interacting with LLMs may give rise to students’ misconceptions. Consequently, identifying disparities in students’ decision making when using AI tools is essential to understand potential negative consequences.

In our work, we examined the field of physics, specifically how students with a strong background in physics interact with ChatGPT to assist them in solving physics questions. We conducted a study with a total of 39 participants with backgrounds in science, technology, engineering and math (STEM) fields from multiple universities. The experimental group (N=27) had unrestricted access to ChatGPT, while the control group (N=12) had access to a search engine only.

Our findings indicate that participants with the CHATGPT condition overly trusted answers generated by ChatGPT. In particular, students often failed to recognize wrong answers given by ChatGPT and largely relied on a copy & paste strategy to solve the posed physics questions. In contrast, participants in the SEARCH ENGINE condition showed higher rates of reflection, as indicated by their sparse use of copy & paste, favoring more thought-through solving strategies.

Our work highlights that there are stark disparities in the interaction behavior between the student groups, provoked by the accessibility of ChatGPT for the experimental group. Even students with advanced domain knowledge struggled to differentiate between correct and incorrect answers given by the LLM and could not use the system effectively. Consequently, there is a need for further research to design AI-based support tools in a way that (1) creates awareness of their inherent uncertainty and (2) allows moderated use that encourages critical thinking.

2. Related Work

The field of language models (LM) offers a variety of possible applications in education. For example, they have been used for multiple-choice question generation [11] or answering [12]. However, since we have to expect students to use LLMs like ChatGPT at home, there is a need to figure out how they utilize these powerful new tools unaided.

Recent advances in natural language processing, initiated by the introduction of the transformer architecture [13], have led to significant progress in the field of language models. The different approaches taken by GPT [14] and BERT [15] models proved to be exceptionally successful. Progress has been steady, with a trend towards increasingly larger models, supported by their scaling laws [16], which suggest that larger size gen-
erally leads to a better model. ChatGPT [1] brought the technology into the public eye, further accelerating the pace of publications and leading to the development of models such as LLaMA [17], GPT-4 [18], and PaLM-E [19]. Some of which even support multi-modal inputs [19]. Language models have shown their potential in many different areas [7,8,9] and are a topic that also influences education [20].

LLMs offer great potential for advancing standard practices and research in education [20]. Several possible applications have been previously suggested, such as personalized learning, lesson planning, assessment and evaluation, to familiarize students with challenges and opportunities of LLMs [20]. Furthermore, a number of studies exist that investigate the use of chatbots based on different technologies in education [21]. The use of chatbots in education offers several advantages, such as serving as a pedagogical tool to help students with disabilities and to help different social groups to close the educational gap that may exist between them [22]. However, none of the systems examined in these works are based on a LLM despite several authors seeing great potential for LLM-based chatbots in the educational domain [23]. It should be noted that LLMs show some weaknesses. Until now, they lack higher-order thinking skills, and their outputs strongly depend on the data they have been trained on, sometimes leading to unreliable outputs [24].

In physics education, there are conflicting reports on the ability of LLMs to solve physics tasks. On the one hand, a few studies have observed inconsistent behavior in ChatGPT’s answers to physics questions [25,26]. These studies showed that ChatGPT often provides incorrect answers to physics questions and concluded that it is unsuitable as a physics tutor or for cheating on homework. Bitzenbauer used this apparent weakness of ChatGPT to foster students’ critical thinking skills by having them generate answers to a question and discuss them critically, leading to an improved perceived usefulness of ChatGPT [24]. On the other hand, other studies demonstrate the strength of ChatGPT 3.5 and 4.0 to solve conceptual multiple-choice questions in physics [27,28]. ChatGPT was able to solve 28 out of 30 items of the force concept inventory correctly [28]. Kieser et al. even found that ChatGPT 4.0 is able to mimic different students’ difficulties when answering conceptual questions, which opens the opportunity for data augmentation, personalized support for students that is sensitive to different difficulties, and support for teachers during task creation [29]. The latter opportunity was studied by Küchemann et al., who compared the characteristics and quality of created physics tasks by prospective physics teachers either using ChatGPT or a textbook. Their findings indicate, that participants who used ChatGPT embedded the tasks less frequently in a real-world context and that most ChatGPT generated tasks were used without modification [30]. These findings point towards the affordances of using ChatGPT in education and the overreliance on AI [31] of participants when using it.

While these articles provide interesting findings and show that using ChatGPT for answering questions present great demands on students, the results were either not verified with real students [25,26] or the problem solving strategies when using ChatGPT were not studied [24,30].

3. Methodology

The related work highlights the existing uncertainties regarding the use of LLMs in general and specifically in the context of physics education. However, to date, little work has
been conducted that allows for moderated and informed usage of such models. We argue that informed usage of generative models is crucial, particularly in educational areas. Our work contributes an investigation specifically into how students interact with LLMs and whether they are aware of their shortcomings. In a mixed-method evaluation conducted online and at two universities (RPTU Kaiserslautern-Landau and LMU Munich), we tasked students with solving given physics problems. Using a between-subject study design, we assessed students’ performance and interaction strategies when having access to different support tools.

As a baseline condition, we had students use an internet search engine (SEARCH ENGINE). This setup represent the de facto standard prior to the advent of LLMs [32,33]. In the CHATGPT condition, students were able to freely use ChatGPT. We recorded the students’ physics knowledge with a pretest (no support tools allowed) and their performance in the main test, as well as inquired about their impressions when interacting with ChatGPT through questionnaires and an exit interview (see Figure 1). Our research was guided by two main research questions:

RQ1: What is the performance of students when being allowed to use ChatGPT in comparison to the students who used a search engine? One main inquiry of our work focused on whether ChatGPT allowed students to perform better when solving the physics questions. We further analyzed the students’ interaction protocols with both tools (SEARCH ENGINE, CHATGPT) to investigate how effectively they used the tool.

RQ2: What are predominant strategies when interacting with ChatGPT compared to search engines? On a meta-level, we were interested in what solving strategies students employed when using ChatGPT and how they differed from the ones used with search engines. From the conducted exit interviews, in combination with the students’ interaction protocols, we distilled predominant strategies when interacting with either tool.

3.1. Physics Question Acquisition

For our main test, we selected four physics questions. To fulfill the requirement that all questions are solvable with school knowledge, we chose questions that require knowledge of six topics of physics taught in school. By choosing four tasks from the International Physics Olympiad [34], a high school competition, we ensured that the tasks were suitable, yet challenging, for university students. All tasks were reviewed by two physics university educators and considered adequate in terms of difficulty and time required for university students. The task texts were adapted in such a way that no picture is necessary for the solution and it was verified that ChatGPT cannot solve the tasks directly and the search engine does not show a page containing the solutions, but both can give hints for obtaining the solution.

We then designed a pretest containing the previously the selected topics. For this, we acquired items from multiple sources [35,36,37] and created our own questions, one of which was inspired by [38].

3.2. Procedure

The study itself was split into multiple parts, as shown in Figure 1. After providing informed consent and an in-depth explanation of the study procedure, the study started
with a self-assessment where participants could rate their physics and ChatGPT knowledge and how often they use ChatGPT (see Section 3.3). Following that, participants had 15 minutes to solve the 17 multiple-choice pretest questions worth 1 point each (max points = 17). Afterwards, participants were allowed to use a modified user interface of ChatGPT or a search engine to help them solve the four physics questions (max points = 12) given a time frame of 30 minutes. The written part of the survey ended with a short questionnaire, including the affinity for technology interaction scale [39] to assess participants’ views on technology, the UMUX-Lite scale [40] to assess usability, and custom questions on perceived accuracy and quality of the tools’ answers, as well as demographics. Throughout the course of the survey, the order of all questions remained unchanged, ensuring the same experience for all participants. For participants attending in person at the university of Kaiserslautern-Landau (N=20), we additionally recorded a short (2-5min) exit interview. After the study, participants were reimbursed with the equivalent of $11 or course credit for a voluntary seminar (N=7). Ethical approval for this study was obtained from the Ethics Committee at the German Research Center for Artificial Intelligence (DFKI).

3.3. Participants

For our baseline condition (SEARCH ENGINE), we acquired 13 participants by providing them the option to do the survey online using university mailing lists from both universities. Of these, 12 participants (Age $\bar{x}$=23.6 y, $s$=2.6 y; 10m, 2n/a; 3 in person, 9 online) fully completed the survey. The students (5 physics, 7 non-physics STEM) were on average in their eighth semester ($\bar{x}$=7.4, $s$=4.3), scored eight points in the pretest ($\bar{x}$=8.2, $s$=3.9, max=12), had an above-average self-reported physics knowledge ($\bar{x}$=62.8, $s$=25.1) coupled with below average experience when using ChatGPT ($\bar{x}$=40.5, $s$=33.4)$^2$.

For the second condition of our study (CHATGPT), we initially recruited 30 participants from two different universities (RPTU Kaiserslautern-Landau, LMU Munich), with a background in physics. They were recruited using mailing lists, posters, and by advertising the study in lectures. Of these 27 participants (Age $\bar{x}$=22.6 y, $s$=4.0 y, 25m, 2f, 27 in person, 0 online) fully completed the survey. Participants were, on average, in their sixth semester ($\bar{x}$=5.3, $s$=3.3). Participants (17 physics, 10 non-physics STEM) scored on average nine points in the pretest ($\bar{x}$=9.2, $s$=3.2, max=15). Using ANOVA, we found a statis-
cally significant difference for the pretest score between physics and non-physics STEM students \((F(1, 37)=11.8, p < .002, \eta^2=.24)\)\(^3\) but not the CHATGPT and SEARCH ENGINE conditions. Further, students reported an above-average perceived physics knowledge \((\bar{x}=58.7, s=18.6)\) and below average experience with ChatGPT \((\bar{x}=42.2, s=24.5)\).

### 3.4. Apparatus

For the CHATGPT condition, we used ChatGPT 3.5 turbo with client side modifications using JavaScript, including a rating scale (good, neutral, bad) to appear with each answer provided by ChatGPT to allow participants to directly voice their opinions. Furthermore, we implemented a download button to be able to save the conversation\(^4\).

For the SEARCH ENGINE condition, we set up a website through which participants could use Google while we were able to collect their search queries.

When students participated in person, we further recorded the ChatGPT conversation log, the participants’ ratings and conducted an exit interview. Additionally, all participants were allowed to use a non-programmable calculator, pen, and paper throughout the study.

### 3.5. Measures

To allow for a holistic picture of how students interact with ChatGPT, we measured student performance through different factors, conducted exit interviews, and analyzed the full student interaction protocols as described in the following section.

#### Student Performance

To evaluate participant performance, a grading schema was created by two physics university educators. Using this schema, two other physicists scored the given answers for the four main questions, independently from each other, awarding between zero and three points per question and participant. We evaluated the inter-rater reliability by calculating the average Cohen’s Kappa \((\kappa=0.72)\) over all main questions, which indicates a substantial reliability [41]. Through discussion both raters reached an agreement in cases where their initial rating differed. The resulting final scores show student performance in answering the main questions. Further, we determined how participants reached their final answers, indicating their problem solving strategy. If the final result of a question was present in the interaction protocol with ChatGPT related to that question, we assigned “extracted from ChatGPT” as strategy. Otherwise, it was counted as “own answer”. Questions that were not answered were counted as “none”. When it was not evident how the answer was obtained, we assigned “random guess” as strategy.

#### Interaction with the support tools

We analyzed the interaction of the participants with their respective tool (ChatGPT or search engine). For the CHATGPT condition, this includes all prompts from participants, respective answers from ChatGPT and associated ratings from participants. Furthermore, for the SEARCH ENGINE search queries were analyzed.

---

\(^3\) Effect sizes are given using \(\eta^2\) (Partial Eta Squared): small (> .01), medium (> .06), large (> .14).

\(^4\) At the time of the study, this feature was not yet available.
Perceived Correctness of ChatGPT Answers  Having two physicists additionally rate all answers given by ChatGPT for correctness enabled us to compare how students rate answers and their actual correctness. With this information, we were able to calculate the false positive rate (FPR), i.e. positively voted answers that are incorrect, and the true positive rate (TPR), i.e. positively rated answers that are correct. In our analysis, we focus on these metrics as they highlight how often information from ChatGPT was assumed to be correct.

Interaction Types  Additionally, we created codes to represent the strategies with which participants created their prompts by categorizing each individual prompt into a coding, comparing and merging them as needed until a consistent representation emerged.

Custom questions  As mentioned in Section 3.2, we administered the ATI [39] and UMUX-Lite [40] as well as two custom questions to inquire about the participants’ impression on ChatGPT correctness accuracy and quality.

Exit Interviews  We conducted exit interviews with 20 participants that were assigned the ChatGPT to further examine qualitative aspects of their interaction. Questions during the interview included asking what strategies were used, how the tool was used and how confident participants were in the correctness of their results.

4. Results

4.1. Student Performance

On average, participants scored $\bar{x}=1.04$ points ($s=1.43$) out of the maximum achievable 12 points in the ChatGPT condition. Most points (nearly 90%) were achieved in questions Q1 and Q3. We found a large positive correlation between the final score and the pretest score, using Kendall’s rank correlation ($\tau=.37$, $p=.02$). No further correlations with respect to the final score were found, in particular for the self-assessed physics knowledge, and study program related demographics such as study subject and semester.

Analyzing how final answers were obtained, we observed that the most prominent strategy was “extracted from ChatGPT” being used in 62% of all cases. Following this, 28% of participants arrived at their “own answer”, 9% of questions were not answered (“none”) and 1% made a “random guess”.

For the Search Engine, participants scored $\bar{x}=1.83$ points ($s=1.27$) on average. Here too, most points (around 95%) were achieved in questions Q1 and Q3. Using Kendall’s rank correlation, we found a statistically significant medium positive correlation ($\tau=.27$, $p=.03$) between the main test score and self-assessed physics knowledge, but none for the pretest score.

Further, we conducted a one-way ANOVA after rank-aligning the data [42] to investigate whether there are significant differences between our two conditions (Search Engine, ChatGPT) with regard to the students’ performances in the main test. We found that students in the Search Engine condition performed significantly better ($F(1,37)=5.5$, $p=.02$, $\eta^2=.13$). Progress in the study program (semester number), course of study (physics, non-physics STEM) and self-rated physics knowledge did not impact the final score as confirmed by ANCOVA tests.
4.2. Interactions with ChatGPT

In total, participants working with ChatGPT created 272 prompts, 165 of which were rated (see Section 3.4). Overall, participants rated 47% of ChatGPT answers positively, indicating that they deemed them to be correct. 29% were rated neutral, and 24% negative, indicating that participants were unsatisfied with them.

Contrarily, our expert physicists only rated approximately 22% as correct, highlighting a mismatch in expectations. This effect is visible throughout all main questions, as depicted in Figure 2. To further analyze intersections in believes of students and experts, we looked at perceived correctness (see Section 3.5). We obtained a high false-positive rate of 57%, i.e., over half of all the answers provided by ChatGPT were believed to be correct by participants but rated incorrect by experts. The true-positive rate of 91%, however, indicates that participants rated most correct answers positive.

![Figure 2](image)

**Figure 2.** The proportion of positively rated ChatGPT answers to students’ prompts visualized for each of the main questions and broken down for students and experts.

**Interaction Types** We identified four main interaction types based on the reviewed ChatGPT interaction logs from all participants: *copy & paste, preprocessing, postprocessing,* and *transformation.* The individual interactions are described in more detail below.

**Copy & Paste** is the most prominent interaction type, where participants transferred the physics question directly to ChatGPT without any changes.

**Preprocessing** is characterized by students trying to reduce the question complexity and using simple priming strategies. They divide a question into multiple parts (P10), ask for formulas (P4), or try to prime the model to improve their results when asking physics questions (P14).

**Postprocessing** builds on already existing answers given by ChatGPT. The participants try to obtain explanations for parts of a question (P1) or correct mistakes they found in the given answer (P12).

**Transformation** is an interaction type where students used ChatGPT to apply some kind of transformation on the data, including translation into another language (P6) and summarizing results (P3).
**Interaction Strategies**  During the study, we noticed that students built their individual strategies to solve the given physics questions based on these interaction types. For example, a participant might start with priming ChatGPT (preprocessing), followed by **copy & pasting** the question and, ultimately, asking for an explanation of some part of the answer (post-processing).

Overall, **copy & paste** was the most used interaction strategy, being used 84 times. **Preprocessing**, the next most common strategy, was used 37 times, followed by **post-processing** (36) and **transformation** (16). In Figure 3a, the distribution of used interaction strategies per question is visualized.

![Distribution of interaction types for each question of the main test for the ChatGPT condition.](image)

**Figure 3.** Distribution of interaction types per question for both conditions. Interaction types that could not be assigned to a question since their content did not include any identifying markers were put into the unspecified category.

### 4.3. Interactions with the Search Engine

To be able to compare how participants of both groups interacted with their respective tool, we describe interaction types and strategies when using the search engine here. The distribution of used interaction strategies is visualized in Figure 3b.

**Interaction Types** We divided the interactions done with the search engine into the same four types as the interactions with ChatGPT (Section 4.2). This allows for easy comparison between the two conditions. There are some minor updates to the interaction types, as some interactions seen when using a search engine were not present when using ChatGPT. **Preprocessing** for the SEARCH ENGINE condition mainly consists of asking for formulas and calculations, while **postprocessing** only encompasses asking for explanations. In the **transformation** interaction, the interaction types “finding answers to related problems” and “trying to find the initial question using keyword search” were added. There were no changes to the **copy & paste** interaction.

**Interaction Strategies** The relations between how often different strategies were used changed considerably from the CHATGPT condition to the SEARCH ENGINE. Here, the most used strategy was **preprocessing** with 64 uses, followed by **transformation** with 17 uses, **postprocessing** (8) and **copy & paste** (3).
4.4. Custom Questions

We calculated the average ATI [39] score of all study participants ($\bar{x}=4.35$, $s=0.79$) showing above-average technical affinity allowing them to adequately interact with the given tools. Additionally, we used the UMUX-Lite [40] questionnaire to calculate a parity score for SUS [43] for the ChatGPT ($\bar{x}=73.05$, $s=9.95$) and the Search Engine condition ($\bar{x}=66.23$, $s=11.62$). Both indicate an above-average system usability. Further, participants rated ChatGPT answers for correctness at $\bar{x}=58.0$ ($s=18.59$) and their quality at $\bar{x}=69.26$ ($s=16.21$) on a visual analog scale from 0 to 100. The search engine answer correctness was rated $\bar{x}=59.6$ ($s=22.8$) and its answer quality $\bar{x}=55.5$ ($s=28.7$). We found no significant differences between the two conditions for all custom questions.

4.5. Exit Interview

We recorded the audio of the ChatGPT exit interviews (59:30 min) and transcribed them using Whisper [44]. To analyze the exit interviews, we used the approach by Blandford et al. [45]. Two researchers coded all interviews separately and merged a final coding tree. From a final discussion, the following themes surfaced: Strategies, Interaction, and Reflection as presented in detail below.

**Strategies** While a diverse set of strategies was employed by the participants, most of them mentioned copy & pasting a question in their exit interview. Different reasons for this were given, such as wanting to see how ChatGPT would deal with the question (P4) or that they did not know how to address the physics question (P7). Other strategies included using ChatGPT like a search engine, e.g., asking for formulas (P1) as it was more convenient. Some strategies indicated a higher level of reflection, such as prior physics problem conceptualization and asking targeted questions (P10). Similarly, ChatGPT was used to explore options for possible solutions and approaches. Here, students identified valuable pieces in ChatGPT answers and showed the ability to detect mistakes and inconsistencies in its argumentation (P4). Though, participants also stated that they had to compromise between speed and correctness of their solutions due to the time constraints. While motivated initially, they tried to offload more work to ChatGPT if time was running out (P1).

**Interaction** When interacting with ChatGPT, participants identified a need to use informed queries. Some tried to achieve this by extracting the most relevant parts of a question from it (P3). Others found that longer texts worked poorly, implying a need for concrete queries to work around this issue (P20) or requiring participants to dig deeper into an answer given by ChatGPT (P12). Interestingly, some participants described their interaction/conversation with ChatGPT as human-like, that the answers looked nice and were very well elaborated (P5). However, selected participants feared that this could delude unaware users (P20).

**Reflection** A number of participants were aware that it is important to reflect on the answers given by ChatGPT, rigorously reviewing them for correctness (P20) and identifying mistakes made by ChatGPT (P1). Especially participants with background knowledge about LLMs were aware of ChatGPT’s weaknesses with regard to physics content and knew what to look out for (P18). Contrarily, for most physics questions, participants showed no sign of actively engaging with the exercises, limiting their reflection (P14).
5. Discussion

Our study provides concrete evidence that students demonstrated vastly different problem solving strategies when having access to ChatGPT and heavily relied on its answers, even struggling to determine their validity. In the following section, we elaborate on these findings and highlight open research questions for the responsible use of LLMs in education.

5.1. Overreliance on ChatGPT Answers Leads to Low Scores

Scored student performance (RQ1) was worse than initially expected (Section 4.1) given our curated selection of exercises. Students using the CHATGPT condition performed significantly worse compared to students in the SEARCH ENGINE condition. Moreover, our study revealed that students in the CHATGPT condition had difficulties detecting if answers generated by ChatGPT were correct or not, as indicated by the high false positive rate of 57%. The unreflected acceptance of presented answers is worrying as it might lead from singular misinformation to general misconception and showcases that there is a definite need to research interactive mechanisms to increase awareness of the uncertainty of LLMs. Contrarily, most search engines are less likely to suffer from this drawback, as presented results are not formulated as definite answers, a design aspect that could potentially inform the design of future interfaces for LLMs.

5.2. Copy & Paste Is The Most Prominent Strategy for ChatGPT Users

This overreliance also manifested when analyzing the employed interaction strategies for the two different tools (Sections 4.2 and 4.3). In the CHATGPT condition, 42% of search prompts are based on copy & paste, highlighting the limited reflection during problem solving. We did observe some participants testing out informed strategies like priming, reducing the question complexity, or correcting ChatGPT (RQ2). However, the low ratio of these informed strategies shows the necessity of teaching users how to use LLMs effectively, allowing them to write prompts to achieve accurate results. The novelty — and thus unfamiliarity — of the interface often enticed users to use the most convenient option (copy & paste) available (RQ2).

5.3. Limitations

Overall, we expected students to score better given the careful curation of our exercises through physics education researchers (see Section 3.1). In hindsight, our questions might have been too difficult for a realistic assessment of how students interact with ChatGPT. However, this result also shows that proper training on how to use LLMs such as ChatGPT might be necessary to achieve good results.

The number of total participants that took part in our study was relatively small. To alleviate this, we made the SEARCH ENGINE condition of the survey available online as well, allowing us to gather more participants. However, due to the online environment, it is possible that the answer quality was lower compared to in person participants. Though, if that were the case, we can assume that the difference between the two conditions would have been even more prominent.
While our recorded results are limited to one specific ChatGPT version (3.5) available at the time of writing, we believe that the implications of this work hold true for future iterations of ChatGPT. Although improved capabilities of ChatGPT can deliver potentially more correct answers, this does not change the fact that students overly trusted its answers and showed little reflection on their assigned tasks, limited learning effects.

5.4. The Potential of Moderated Use of LLMs

Our analysis revealed a need to think about the design of educational systems that use LLMs. We need to moderate interaction with language models such as ChatGPT in a way that students can profit from the vast abilities of such tools while simultaneously reducing the negative impact it can have on the students’ learning progress. Informed use can be achieved by raising awareness of LLM caveats and educate users on how to best use them. Though we argue that, to leverage the full potential of these models, we should strive to achieve moderated use: a usage that allows students to interact with ChatGPT as a guidance teacher or sparing partner to formulate and explore ideas to solve a physics problem. Such a system should carefully guide students towards the solution, introducing necessary concepts but allowing critical thinking and reflection while still being enjoyable and effective to use. If we can demonstrate the benefits of moderated LLMs compared to unrestricted LLMs to students in terms of their ability to learn and progress, we can certainly change and evolve the current ways of teaching. A possible way to moderate LLMs would be to change their output behaviour using prompt engineering as we have done in a different work [46].

6. Conclusion

In this work, we analyzed the impact of ChatGPT on problem solving strategies of students. We found that students who used ChatGPT performed significantly worse compared to those using a search engine. Furthermore, stark differences in user interaction manifested, where ChatGPT users mainly relied on copy & pasting questions and answers, while search engine users used more refined strategies such as searching for formulas. This highlights missing reflection and limited critical thinking as two of the main issues when using LLMs in education. To combat this, we — first and foremost — suggest to inform students more adequately of the shortcomings of these models. Though ultimately, we want to converge towards moderated LLMs, specifically designed to support students in a meaningful way by encouraging critical thinking.

Acknowledgements

This research is supported by the European Union’s Horizon 2020 Programme under ERCEA grant no. 952026 HumanE-AI-Net and by the German Federal Ministry of Education and Research (BMBF) through the project KI4TUK.
References

[41] Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977:159-74.
“Computer, Generate!” – Investigating User-Controlled Generation of Immersive Virtual Environments

Carina LIEBERS, Niklas PFÜTZENREUTER, Jonas AUDA, Uwe GRUENEFELD, and Stefan SCHNEEGASS

University of Duisburg-Essen, Schuetzenbahn 70, 45127 Essen, Germany
GenerIO.ai, Schuetzenbahn 70, 45127 Essen, Germany

Abstract. For immersive experiences such as virtual reality, explorable worlds are often fundamental. Generative artificial intelligence looks promising to accelerate the creation of such environments. However, it remains unclear how existing interaction modalities can support user-centered world generation and how users remain in control of the process. Thus, in this paper, we present a virtual reality application to generate virtual environments and compare three common interaction modalities (voice, controller, and hands) in a pre-study ($N = 18$), revealing a combination of initial voice input and continued controller manipulation as best suitable. We then investigate three levels of process control (all-at-once, creation-before-manipulation, and step-by-step) in a user study ($N = 27$). Our results show that although all-at-once reduced the number of object manipulations, participants felt more in control when using the step-by-step approach.

Keywords. Generative AI, Virtual Reality, Human-Controlled Scene Generation

1. Introduction

The Holodeck from Star Trek© enables users to generate realistic environments by expressing their desires with simple voice commands next to manual controls. As artificial intelligence (AI) progresses, generating individualized virtual worlds becomes more feasible, as virtual reality (VR) allows users to create and experience such worlds. However, individualized environments are too complex to describe with short voice commands. The question remains of how to give humans control over the generation.

To empower users to generate virtual worlds, we must understand the extent to which they need control during the generation process. User-controlled generation includes many aspects, from object creation to adjusting sizes, textures, and positions to broader environmental elements. To reduce complexity, we break the generation down into individual objects (i.e., the building blocks of the environment). Nevertheless, users still have to manage object creation, type, placement, aesthetics, animation, and interactivity. Although in-depth customization offers great autonomy, it may also be inefficient. Thus, the question remains of how different control flows affect efficiency and user experience.

Although existing literature explored user-driven virtual environment design, it remains unclear which control flow enhances user’s control. Prior works investigated gen-

1Corresponding Author: Carina Liebers, carina.liebers@uni-due.de
eration through object catalog selection [1,2,3] and adjustments of spatial [4,2] and appearance [4,3] attributes, but these often do not generalize well. Furthermore, none assessed user control for generating virtual environments.

To investigate human control in virtual environment generation, we designed a VR testbed, including the input modalities (controller, hand, or voice input) and generation steps (creation, movement, scaling, coloring, and deletion). In a pre-study, we assess which input modalities support the generation steps best, revealing a combination of initial voice and subsequent controller input. Using these modalities, we examine how humans can control the generation process to meet their expectations, using the control strategies all-at-once, creation-before-manipulation, and step-by-step (RQ1). While all-at-once reduced object manipulations, participants felt higher control using step-by-step.

Contribution Statement. We propose a VR test application for environment generation, including controller, hand, and voice input for the generation steps. Using the most suitable input modality for each generation step from a pre-study ($N = 18$), we conduct a user study ($N = 27$) to compare three control flows (all-at-once, creation-before-manipulation, and step-by-step), assessing user control during generation.

2. Related Work

Since we aim to assess the optimal control flow for generating virtual environments within VR, our research is grounded in input modalities and scene-generation methods for VR applications.

2.1. Object Interaction Techniques

To create a virtual world, for instance for level design, users need to interact with the system and the generated objects. These interactions encompass object manipulations, such as generating and deleting objects, and spatial actions involving positioning and scaling objects.

Since gestures and speech constitute language [5], approaches like “Put-That-There” [6] combine gesture and speech to spatially manipulate objects. Participants particularly preferred speech commands such as “move” for translation, but gestures for rotation tasks [7]. For object translation and rotation in VR, research has shown that hand-held controllers tend to offer higher performance, usability, and user preference compared to hand-tracking solutions [8]. Further, 3D interfaces performed better than speech interfaces for positioning and rotating during furnishing tasks in VR [9]. In the context of scaling operations, gesture-speech-based interactions are more effective than grasping interactions [10]. Specifically, when gesture and speech were compared, gestures were rated as the superior modality for scaling objects [7]. For object creation and deletion as well as text input, speech commands have been found to outperform all other input modalities [7,9]. Thus, the input modality depends on the specific tasks [11,12].

Controller-based or gesture-based interactions are typically preferred for translation and rotation tasks, while speech is often preferred for non-spatial operations. However, it is important to note that many of these studies primarily focused on Augmented Reality (AR) environments. As our work involves creating subjective virtual environments that
encompass both spatial and non-spatial operations, we conducted a pre-study to evaluate which input modalities are best suited for our specific use case in VR.

2.2. User-Assisted Scene Generation Using Machine Learning

Machine Learning (ML) methods have been widely used for the creation of virtual scenes and objects in VR and AR environments. These methods typically utilize images or videos as input to generate 3D models and scenes [13]. In the following, we provide an overview of related research that focuses on user-assisted virtual scene and object generation, and ways in which users can interact with generative methods.

To create a virtual world within VR, users primarily select objects to design the environment. They can create a world through catalog selection [1,2,3] or scanning physical environments [14,15]. Environment building methods vary. Some allow direct object placement via controllers [14] or combine voice and hand input for placement [4]. Others first modify the terrain before object manipulation utilizing hybrid wand/tablet interface [16]. VR Safari Park [17] offers a novel approach, letting users add virtual blocks to a world tree, representing entities, supporting especially novices with an effective overview. Environment generation from text input starts with extracting the objects via voice [18] or text input [19,20,21,22]. To translate requests into actual scenes, the objects are used to create a semantic scene graph to generate sub-scenes from 3D databases, which are later adjusted to match user intentions [19]. Alternatively, scenes can be formed by considering inferred object relations [21,20], scanning real-world environments [15,23], or comparing input against annotated datasets and reference images [22].

To generate 3D models for building virtual scenes, current approaches compute multiple images of an object. Utilizing AI systems, especially GANs (Generative Adversarial Networks), facilitates the generation of images from a textual description. They excel in generating images using unlabeled training data [24], yet diffusion models surpassed them in deriving images from text, offering better scene complexity and image quality [25,26]. In previous works, users could control the object generation using voice input, adjusting attributes like style, shape, and categories of the objects generated using pre-trained text-to-image model [27]. User-driven control of ML techniques has primarily focused on image generation. One example is GANzilla [28], which lets users highlight areas of interest, analyze clusters, and explore alternate image generation outcomes. In this work, we aim to assess user control strategies for generating a virtual environment through generating 3D objects using the Point-E generative network [29].

3. Generating Immersive Environments

To investigate how to give users control over virtual environment generation, we first define the scope of our generation procedure. Then, we describe the selected input modalities and control flows and our VR environment as a testbed for the evaluation.

3.1. Overview of the Generation Procedure

Immersive environments consist of numerous elements, from scene-specific settings such as lighting and weather conditions to individual objects’ details and appearances [1]. To reduce the complexity of the generation, we perceive immersive environments as a
symphony of individual objects existing in the environment. We consider general scene aspects by using the example of a virtual garden with a fixed configuration (i.e., preset lighting and weather conditions). For the generation process of the objects, we identified a set of five different generation steps that can be used to create or manipulate an object. In the first step, the object itself is created in step creation. Afterwards, the objects can be manipulated in a spatial manner resulting in the steps: movement and scaling. Last, the appearance of an object can be changed in the steps coloring and deletion.

3.2. Input Modalities

Existing literature suggests controller is effective for Grab-and-Place tasks [8], hand has lower accuracy for precise tasks [8,30] but being more intuitive for users [31,30] and voice input offering easy usage for simple command-based tasks [9] while being used in more application cases [32]. Given the variety of tasks during environment generation, it remains unclear which input modality facilitates the control process best: controller input, hand input, and voice input.

Controller Input. To create an object, we use a virtual keyboard in VR, motivated by Weiß et al. [9]. Users can then translate, rotate, or scale the object either by pointing at or grabbing the object and performing the wished transformation, as suggested by Whitlock et al. [33]. For coloring, we employ a spray gun metaphor, and deletion is achieved by grabbing the object and pressing a button.

Hand Input. The hand input is designed similarly to controller interaction. Users create objects using a keyboard, move, and scale objects by grabbing them, color objects using a spray gun, and delete objects by pressing a button. To select a letter, grab an object, activate the spray gun, or press buttons in the graphical user interface (GUI), users perform the pinch gesture by touching the thumb with the index finger.

Voice Input. Voice input allows speaking a specific command, which is transferred into the desired action. We mapped each generation step to keywords to improve request recognition when using natural language. Following Bolt’s method [6], we integrated voice commands with pointing gestures to select the intended object or specify a location.

3.3. Control Flows

A crucial aspect is understanding what temporal order (referred to as control flow) users want to generate the immersive worlds. We define three control flows: all-at-once, creation-before-manipulation, and step-by-step. While the Holodeck enables environment generation with a single voice command (realized as all-at-once), an alternative is to individually create each object and modify it directly after creating (referred to as step-by-step). We further introduce a mixed method (creation-before-manipulation), combining both. We aim to assess:

RQ1: Which control flow for generating virtual environments supports users best to achieve results that fulfill their goal?

All-At-Once. A detailed initial voice input allows the generation of the entire environment. The request includes all objects and their attributes like location, rotation, scale, and appearance. Users can make manual post-generation adjustments.
Creation-Before-Manipulation. The mixed control flow involves initially requesting all environment objects at once, with subsequent adjustments made in separate steps. It allows pausing object creation, offering users greater flexibility.

Step-By-Step. This approach involves creating and manipulating each object one at a time. Users first generate a single object through a distinct request and then apply spatial and appearance adjustments, similar to De Leon et al. [1].

3.4. Implementation of the VR Testbed

Following, we describe the developed VR environment and the input modalities. We used Unity3D [34] for deployment on the Meta Quest 2 [35]. To realize the controller and hand input, we used the Meta Interaction SDK [36], and for voice input, we used the Wit.ai speech-to-text interface [37]. Next, we describe the VR environment as well as the applied generation and processing technologies.

The Virtual Garden Environment. We designed a garden environment next to a white house, surrounded by a wooden fence measuring 4x4 meters. To enhance realism, trees are positioned outside of the garden under a mostly sunny sky. We include two interaction panels hanging from the veranda: one panel displays the ongoing task and its duration, and provides a button to confirm task finalization. An additional interaction panel is present for controller input and hand input.

Voice Recognition. In voice mode, users could activate voice commands using the controller’s B button. The system recorded voice until a decreased volume signaled speech completion, or for a maximum of 20 seconds. All recordings are visualized as feedback. Given the possibility of non-native English speakers participating, recordings were translated to English using the deep-translator package [38] in Python. Afterward, Wit.ai’s speech-to-text transcribed the speech, determining its intent and identifying entities like objects or colors. A command executes if recognized with over 80% confidence. A successful recognition triggered audio feedback.

Object Generation. Object requests are processed by the Point-E generative network [29], which provides point clouds of the objects integrated using REST. We generate meshes from these point clouds for colliders to ensure accurate object interaction. The point cloud and its mesh are loaded into the garden during runtime. With that, we enable users to create every imaginable object without being limited to an existing database.

4. Pre-study: Input Modalities for Environment Generation

Given the variety of tasks involved in generating environments, it’s unclear which input modality best facilitates the control process, as related work primarily focuses on specific tasks [12,11] and AR environments. Thus, in our pre-study, we aim to identify the optimal input modality for the generation steps in VR. We compare the independent variables input modalities with three levels: controller input, hand input, and voice input, and tasks groups with three levels: creation, spatial manipulations (movement and scaling), and appearance manipulations (coloring and deletion). This research sought to enhance understanding of VR interaction design, particularly which input modalities best support user engagement and task efficiency in a virtual environment.

We conducted a controlled, within-subjects laboratory study to explore the suitability of three input modalities for the different task groups (see Figure 1). During the study, our participants should generate objects of varying complexity (creation), move and scale
objects to predetermined specifications (spatial manipulation), change object colors, and delete them (appearance modification) with all input modalities. The study incorporated a Latin square design to counterbalance the input modality and task group order. We measured the following dependent variables: task completion time (TCT), spatial accuracy, usability, user experience, perceived TCT, expectation conformity and the preferred modality, alongside qualitative feedback through semi-structured interviews after each task group. Details of the variables are outlined in Table 1.

To determine the required participant number, we conducted an a priori power analysis using G*Power for a repeated-measures within-factors analysis of variance (ANOVA), given $\alpha = .05$ and effect size $f = .4$. Following our experimental study design, we set the number of groups to 5 and the number of measurements to 5. G*Power 3.1.9.7 suggested a total sample size of 15 ($\lambda = 24, F = 2.61$). Due to our randomization, we invited 18 participants for our study. Of the volunteers (15 identified as male, 3 identified as female, and 0 as non-binary), all except one were right-handed, whose median age was 26.5 years ($SD = 4.34, Min = 19, Max = 35$). In terms of VR experience, 1 participant reported to have “never” used VR, 5 participants used VR “once a year”, 4 “once a month”, 7 “once a week”, and 1 participant used VR “daily”. For voice control experience, 2 participants reported to have “never” used voice control, 5 participants used Voice “once a year”, 6 “once a month”, 1 “once a week”, and 4 participants used voice control “daily”. Conducted in a room with a $4m \times 4m$ open space, we ensured a safe and controlled setting for the VR interaction. To ensure the participants’ privacy, we only recorded pseudonymized data. Our local ethics committee approved the study.

**Voice Input and Controller Input Are Preferred for Creation** Our quantitative data shows no significant difference in TCT among the input modalities for creation. Despite this, participants perceived hand input as slower than voice input (Perceived TCT, $t(4.66) = -3.42, p = .005 < .01$), especially with virtual keyboards. Regarding usability, voice input significantly outperformed hand input ($t(4.99) = -4.10, p < .001$), whereas participants often referred to the hand input’s accuracy issues as drawbacks.

<table>
<thead>
<tr>
<th>Generation Steps</th>
<th>Creation</th>
<th>Spatial Manipulations</th>
<th>Appearance Manipulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Modality</td>
<td></td>
<td>Movement</td>
<td>Scaling</td>
</tr>
<tr>
<td>Controller</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hand</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voice</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.** Tasks in the Pre-Study: We compared the input modalities (controller, hand, and voice) for the generation steps (creation, movement, scaling, coloring, and deletion), categorized into three task groups: creation, spatial manipulation, and appearance manipulation.
Voice input was praised for bypassing manual input despite occasional inaccuracies. Participants stated that controller interaction was more cumbersome than a voice query for creation, although they enabled faster typing. Voice input also surpassed hand input in pragmatic user experience ($t(5.33) = -2.96, p = .017 < .05$) and outperformed controller input in the overall scores ($t(5.33) = -2.58, p = .041 < .05$). Our participants’ modality rankings mirrored these insights, favoring controller input ($t(4.67) = -3.43, p = .005 < .01$) and voice input ($t(4.67) = 4.93, p < .001$) over hand input for creation between the input modalities. These trends indicate a preference for voice and controller input during creation tasks.

Controller Input Outperforms Hand and Voice Interaction for Spatial Manipulations

For spatial manipulations, voice input was less accurate than controller input ($t(3.8) = -6.08, p < .001$) and hand input ($t(3.8) = -3.48, p = .004 < .01$) for translation and rotation, and many participants found precise modifications challenging. In expectation conformity, controller input led to significantly more expected outcomes than voice input ($t(3.01) = 9.33, p < .001$). The usability scores display higher usability for controller input than voice input ($t(4.22) = 5.62, p < .001$). Comparing hand input and voice input, hand input received higher pragmatic User Experience Questionnaire (UEQ) ($t(4.01) = 3.12, p = .011 < .05$) and preferred modality scores ($t(3.55) = -5.36, p < .001$). While participants indicated technological limitations for voice and hand inputs in the interviews, hand input was preferred for intuitive interaction. In summary, controller input displayed an advantage for spatial manipulations, as evidenced by qualitative feedback and higher accuracy in translation and rotation. It outperformed hand ($t(4.22) = 3.22, p = .009 < .01$) and voice inputs ($t(4.22) = 5.62, p < .001$) in System Usability Scale (SUS) and pragmatic UEQ scores (hand input: $t(4.01) = 3.39, p = .005 < .01$, voice input: $t(4.01) = 6.5, p < .001$).

Controller Input and Voice Input Are Both Suitable for Appearance Manipulations

Contrary to our expectations, controller input was significantly faster than hand input ($t(9.8) = -10.25, p < .001$) and voice input ($t(9.8) = -7.88, p < .001$) in appearance manipulations. Although we found no significant difference in TCT between hand input and voice input, participants perceived hand input as slower than voice input (perceived TCT). Hand input further received lower usability scores than controller input ($t(3.87) = 5.11, p < .001$) and voice input ($t(3.87) = -6.03, p < .001$). Participants faced challenges with the pinch gesture interfering with the appearance manipulations for hand input. Voice input ($t(4.14) = -4.94, p < .001$) and controller input ($t(4.14) = 4.22, p < .001$) outperformed hand input in user experience and were strongly favored (preferred modality, controller input: $t(3.88) = -5.67, p < .001$, voice input: $t(3.88) = 6.70, p < .001$). Between them, controller input excelled in efficiency (TCT), while voice input was also highly valued for its user experience.

5. Human Control in the Generation of Virtual Environments

We assess which control flow for generating virtual environments provides the most control over the generation to fulfill their initial vision (RQ1), focusing on the temporal user engagement in the generation process, ranging from an early to late user intervention. Following, we outline the study details and its findings. We used the same apparatus as in the pre-study (see Section 4) and obtained approval from our local ethics committee.
Table 1. Variables and Their Measurement for Evaluating Input Modalities in the Pre-Study.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Calculation</th>
<th>Measurement</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task completion time (TCT)</td>
<td>Time of first to last task-related action</td>
<td>Perceived TCT</td>
<td>7-Point Likert items: “I was able”</td>
</tr>
<tr>
<td>Spatial Accuracy</td>
<td>Translation (positional offset), rotation (angular distance), scale (scale vector distance)</td>
<td>Expectation Conformity</td>
<td>7-Point Likert items: “The outcome was as expected.”</td>
</tr>
<tr>
<td>Usability</td>
<td>System Usability Scale (SUS) questionnaire [39]</td>
<td>Preferred Modality</td>
<td>Modality rating from best to worst</td>
</tr>
<tr>
<td></td>
<td></td>
<td>User Experience</td>
<td>User Experience Questionnaire (UEQ) Short [40]</td>
</tr>
</tbody>
</table>

5.1. Study Design

We conducted a within-subject design, examining the independent variable control flow, with three characteristics: all-at-once, creation-before-manipulation, and step-by-step. We hypothesized that a more automated process (e.g. all-at-once) would likely diminish the interactions users need to shape their environment. Fewer interactions could lead to higher usability and allow users to realize their visions faster. Additionally, more interactions could lead to prolonged adjustment times, potentially resulting in users feeling fatigued, bored, or stressed, potentially negatively affecting their user experience during the generation. Moreover, given the extended adjustment period and possible negative feelings, we assumed a less automated approach (e.g. step-by-step) would result in a less satisfying scene outcome. Thus, we hypothesized:

H1.1 All-at-once leads to higher usability than step-by-step.
H1.2 Step-by-step leads to a worse user experience compared to all-at-once.
H1.3 Step-by-step leads to a less satisfying scene outcome compared to all-at-once.

During the study, participants generated a virtual garden with five objects based on three scene templates to negate the influence of object types, colors, and positions. We adopted a Latin square design to counterbalance scene task and control flow, leading to nine configurations. Our measured dependent variables can be found in Table 2. Following the study tasks, we conducted semi-structured interviews for qualitative feedback.

5.2. G*Power Analysis

To determine our required sample size, we conducted an a priori power analysis for a repeated-measures within-factors ANOVA, given $\alpha = .05$ and effect size $f = .4$. Following our study design, we set the number of groups to 3 and measurements to 3. G*Power 3.1.9.7 suggested a total sample size of 18 ($\lambda = 17.28, F = 3.32$), which we exceeded by half due to non-existing references for similar studies.

5.3. Participants

We recruited 27 participants (21 male, 6 female, 0 non-binary) for our study of which 17 also took part in the pre-study. All except one were right-handed, and 28.0 years old in the median ($SD = 3.97, Min = 19, Max = 35$). The participant self-reported to have used
VR 3 times “never”, 6 times “once a year”, 7 times “once a month”, 10 times “once a week”, and 1 times “daily”. They stated to have used voice control, 3 times “never”, 12 times “once a year”, 5 times “once a month”, 3 times “once a week”, and 4 times “daily”.

5.4. Procedure

Before the study, we briefed participants about the goals and procedure, emphasizing their right to withdraw at any time, and informed them of the recording of the TCT and task correctness. After they provided written consent, the study began. During the study, the participants generated a virtual garden based on three provided templates (see Figure 2), which included pre-defined objects with distinct positions and colors, using the control flows. They first entered a training scene for each control flow to familiarize themselves with the generation process. Afterward, participants moved to the actual task, followed by filling in the SUS, UEQ, and an individual questionnaire before progressing to the next control flow. The study concluded with a semi-structured interview to capture qualitative feedback. An entire session took 45 minutes per participant.

5.5. Study Tasks

We utilized the best input modalities for the generation steps from the pre-study: voice for creation and controller input for the subsequent spatial and appearance adjustments.

**Step-By-Step.** Participants use voice input to create one object at a time and then adjust its color and position with the controller. The process halts after each creation to prevent simultaneous creations until the next object is initiated.

**Creation-Before-Manipulation.** Participants use a single voice command to initiate multiple object creations, which appear one after another. They can pause the process to adjust an object’s appearance or position, allowing step-wise adjustment of all objects.

**All-At-Once.** Participants generate the entire environment with one voice command, including objects, colors, and positions. The generation process remains uninterrupted until completion. Afterward, participants can make manual adjustments.

<table>
<thead>
<tr>
<th>Variables and Their Measurement for Evaluating Control Flows</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table 2</strong>. <strong>Table 2. Variables and Their Measurement for Evaluating Control Flows</strong></td>
</tr>
<tr>
<td><strong>Measurement</strong></td>
</tr>
<tr>
<td>TCT</td>
</tr>
<tr>
<td>Task Correctness</td>
</tr>
<tr>
<td>Manipulation Time</td>
</tr>
<tr>
<td>Manipulation Number</td>
</tr>
<tr>
<td>Usability</td>
</tr>
<tr>
<td>User Experience</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
5.6. Results

Quantitative Analysis  For each variable, we first tested for homogeneity and normality. If met, we proceeded with an ANOVA for parametric data, else we first performed an Aligned Rank Transform (ART). Following, we only report significant results from post hoc tests with Bonferroni correction.

We measured a manipulation number of 13 (IQR=16.5) in all-at-once, 18 (IQR=16.0) in creation-before-manipulation, and 24 (IQR=10.0) in step-by-step. We investigate the effect of control flow on the manipulation number of our participants, and found a significant difference with a one-way ANOVA ($F(2, 52) = 10.10, p = .001, \eta^2_p = .28$). We found a significant difference for spatial manipulation between step-by-step and all-at-once ($t(3.88) = -4.47, p < .001$), and between creation-before-manipulation and all-at-once ($t(3.88) = -2.64, p = .033 < .05$). We can conclude that all-at-once leads to a lower manipulation number than the step-by-step and creation-before-manipulation.

Our participants rated the perceived control of all-at-once with 5 (IQR=3.0), creation-before-manipulation with 6 (IQR=1.5), and step-by-step with 6 (IQR=1.5). With one-way ANOVA, we found a significant difference of control flow ($F(2, 52) = 3.88, p = .027 < .05, \eta^2_p = .13$) on the participants’ perceived control of the generation. Post hoc analysis revealed a significant difference between step-by-step and all-at-once ($t(5.81) = -2.58, p = .038 < .05$). Thus, we can conclude that our participants perceived higher control using step-by-step than using all-at-once.

Qualitative Analysis  Using the thematic analysis [41], two researchers analyzed the interview responses. We iteratively refined codes grouped them into topics and identified four main themes: Application Design and User Experience, Advantages/Disadvantages Input Modalities, Differences of Control Flows, and Use Cases of Generation.

Advantages/Disadvantages Input Modalities. Many participants (15) found voice input effective. However, challenges, such as the need for clear or loud statements without pausing during voice commands, often resulted in aborted or incorrect outcomes. A few noted task complexity, mentioning difficulties with longer keywords for positioning.
For the controller, feedback focused on its ease of use and button allocation. 6 participants found the spray gun easy to use, while 9 felt manual object positioning intuitive.

**Application Design and User Experience.** Many participants (10) enjoyed the application and found the VR environment creation engaging. "The system was great to use. You felt like you could do whatever you wanted" (P21). However, participants reported technical issues, especially with voice input, including objects not spawning or being adjusted as requested. They further had difficulties in selecting objects that were closely positioned with controller input. 6 participants wished for a different control flow, suggesting creating single objects with their appearance attributes using voice input followed by manual spatial manipulation.

**Control Flow Differences.** Although appreciating the idea, our participants expressed all-at-once required a high mental effort due to submitting long sentences. It required envisioning the entire scene before making the request, including object names, colors, and positions. Despite these challenges, most participants found all-at-once efficient. They felt it sped up the generation process by setting colors and placements in advance, reducing their overall effort, "It was definitely the most efficient and easiest to use." (P21, All-at-once). Many participants (9) praised its innovation and expressed excitement when their requests were executed accurately. For step-by-step, several participants (7) found it less convenient, more time intensive, and less effective due to the manual adjustments. Although they perceived a higher workload for step-by-step, many participants (10) found it more reliable and predictable, resulting in higher felt control. One stated having a better overview of what he was doing, while another emphasized fewer errors. Some viewed creation-before-manipulation as having a higher workload. Others found it more efficient, facilitating adjustments while awaiting object generation, "I could already say several things, I already let it work and continued with the next object or checked another" (P19, Creation-before-manipulation). Due to making manual adjustments, a few participants perceived greater control, similar to step-by-step.

**Use Cases of Generation.** Discussing potential applications, many participants envisioned home-related uses like interior or garden design (19). Architectural, urban planning, and video game scene designs were also suggested, with several referencing the Sims™ game. Overall, participants saw potential in all control flows. For tasks demanding precision, they favored step-by-step for its control. Creation-before-manipulation was seen as efficient, saving time while allowing subsequent adjustments. For scenarios with time constraints, where the primary goal is not creation, they considered all-at-once for its time-saving and streamlined process.

### 5.7. Discussion

Following, we discuss the findings of our assessment. While the progress of generative AI might impact input modality selection, generation speed, and visual results, we consider control flows of virtual environment generation as largely unaffected.

**All-At-Once: Efficiency in Time-Sensitive Tasks** We hypothesized all-at-once to enhance usability by reducing user interactions, based on literature [42,43]. However, our findings did not support H1.1, though they did highlight its efficiency in reducing manipulations. Participants found manual adjustments in other methods lead to an increased workload, as supported by Kieras et al. [44], and appreciated all-at-once when time is prioritized over detail. Despite fewer manipulations, user experience scores did not in-
crease, potentially due to voice input’s limitations and the mental effort of verbalizing concepts, supporting prior studies [6,45,7,9]. With no significant differences in usability, TCT, and task correctness across control flows, all seem suitable for generating virtual environments. Still, all-at-once’s efficiency might benefit quick design iterations, like indoor or architectural planning, enabling faster testing of various designs.

**Step-By-Step: Preferred for Fine-Granular Tasks**  We hypothesized that increased interactions in step-by-step might impair user experience by adding cognitive load [44] and extending adjustment times, negatively impacting user experience [43,42] and scene quality. Contrary, our findings did not support **H1.2** and **H1.3**. However, participants found that step-by-step offered enhanced control compared to all-at-once, particularly for precise spatial tasks, and noted reduced mental workload through iterative adjustments. Thus, all-at-once’s drawbacks, particularly the cognitive demands and limited voice input affordance, reduced perceived control, in line with Myers et al. [46]. Further, participants preferred step-by-step for its ease of use and tactile feedback from controllers, aiding in precise object placement. Thus, it might be especially suitable for precise tasks where control and accuracy are primary, while time is less critical, such as fine-grained customization of gaming environments or room configurations for marketing.

**Control Flow Variants**  Participants suggested alternative control flows, like combining voice input for object creation and controller input for spatial adjustments. Further, they wished for customizable processes based on task and personal preferences, in line with related work supporting adjustments of user interfaces to user preferences [47,48]. When used in adaptive systems, personal preferences could be recognized after some time and integrated directly into personal settings regarding the control flow. This finding further suggests exploring other mixed control flow combinations to enhance user satisfaction and efficiency while facilitating user adjustments to support their individual needs.

6. **Conclusion**

We investigated how humans can control the generation of virtual immersive environments in VR by assessing three control strategies (all-at-once, creation-before-manipulation, or step-by-step). To do so, we first designed a VR testbed. In a preliminary study, we utilized it to evaluate the input modalities (controller, hand, and voice input) for different generation tasks. Combining voice for object creation and controller for spatial and appearance manipulation was the most effective. Thus, we employed voice for creation and controller for spatial and appearance manipulation to evaluate the three control flows regarding their level of control over the generation process. We found that generating the environment all-at-once resulted in fewer object manipulations, which points to enhanced efficiency – a sentiment echoed by our participants. However, participants felt more control over the generation process when using step-by-step because the cognitive demands at each step were manageable, facilitating more fine-grain adjustments with fewer errors. Our findings suggest that the control strategy should consider the application’s focus and the user’s preference for efficiency or control.

**Acknowledgements.** This work is funded by the German Federal Ministry of Education and Research (01IS21068B).

**Declaration of Conflicting Interests.** The authors declare no conflicts of interest.
References


Human-AI Engineering for Adults

André MEYER-VITALI a,1, and Wico MULDER b

aDFKI, Saarbrücken
bTNO, Groningen

ORCiD ID: André Meyer-Vitali https://orcid.org/0000-0002-5242-1443, Wico Mulder https://orcid.org/0000-0002-8607-0055

Abstract. The engineering of reliable and trustworthy AI systems needs to mature. While facing unprecedented challenges, there is much to be learned from other engineering disciplines. We focus on the five pillars of (i) Models & Explanations, (ii) Causality & Grounding, (iii) Modularity & Compositionality, (iv) Human Agency & Oversight, and (v) Maturity Models. Based on these pillars, a new AI engineering discipline might emerge, which we aim to support using corresponding methods and tools for ‘Trust by Design’. A use case concerning mobility and energy consumption in an urban context is discussed.

Keywords. Software Engineering, Artificial Intelligence, Models, Modules, Trust, Causality, Robustness, Explainability, Agency, Multi-Agent Systems, Smart Cities, Human-Computer Interaction, Context-Aware Pervasive Systems, Maturity Model

1. Introduction

The current wave of Artificial Intelligence (AI) is characterised by Deep Learning [1, 2], Transformers [3, 4] and Large Foundation Models [5]. Whilst the impact of such systems touches almost all veins of our society, it seems that we are reaching the limits of controlled engineering of these large, highly interconnected, AI-based systems.

On the one hand, we see their complexity increase on an individual level, as well as on their connected dependency levels, whilst on the other hand, we see a growing lack of experience on the level of their design and engineering. The complexity of existing AI models is often beyond our understanding, and the methods and processes to ensure safety, reliability, and transparency are lacking. This poses serious risks at the level of trustworthiness, particularly when it comes to critical applications with significant physical, economic, or social impact. The AI systems used in such applications are required – for example by the European AI Act – to have been thoroughly designed, validated and certified according to well-defined criteria.

Recent developments in Generative AI are based on so-called ‘Foundation Models’, which can appear as Large Language Models (LLM) or as similar multi-modal models of still images, videos and others. The transformer architectures that generate these
models convert huge amounts of text or other media content into statistical models of co-occurrence of tokens (parts of words or other features).

Many are not aware that deep learning does not support a real understanding of problems. At a first glance, these models for generative AI seem to understand human language and creative expression. However, as they are uniquely based on producing probabilistic assemblies of tokens, they do not even understand language itself. There is no grammar involved or any form of semantics. They only reflect high-dimensional statistical correlations.

Great disillusionment set in as problems, such as insufficient internal representation of meaning (interpretability and transparency), susceptibility to changes in the input signal (robustness), lack of transferability to cases not covered by the data (generalisation) and, last but not least, the thirst for big data and processing itself (efficiency, adequacy, sustainability), became apparent. Some of these problems are a direct result of the massive use of deep learning black-box methods that depend solely on data [6].

To increase the grip and understanding of the outcomes of large neural models, new approaches combine data-oriented machine learning with symbolic conclusions and the explicit representation of knowledge [7, 8].

Such types of approaches are being advanced by the term ‘Trusted AI’. Trusted AI aims to create a new generation of AI systems that guarantee functionality, allowing use even in critical applications [9]. Developers, domain experts, users, and regulators can rely on performance and reliability even for complex socio-technical systems. Trusted AI is characterised by a high degree of robustness, transparency, fairness, and verifiability, where the functionality of existing systems is in no way compromised, but actually enhanced.

Foundation models are not trustworthy, because they lack any kind of understanding of truth, facts, time, space, concepts, reasons, causes and effects. As they are not consistent, transparent, robust and reliable, it is very risky to trust them in critical applications. Even when they seem to give reasonable answers from time to time, it is impossible to predict when they will fail and start to hallucinate.

We need to stop reinventing the wheel; learning from scratch, but understanding nothing. Instead, we need to use existing knowledge, build on experiments and experience, formulate and validate new hypotheses and theories, in order to gain knowledge and insight at a higher level, and to explain why events happen, predictions are made and decisions or actions are taken. This requires a reinforced attention for engineering processes with an aim to improve scientific progress where one can stand on the shoulders of giants.

This paper therefore sets out five pillars of AI Engineering. Together, they form a supportive framework that fosters ‘Trusted AI by Design’. Section 2 discusses the five pillars and section 3 provides a use case in a smart city context.

2. Trusted AI Engineering

There is a dilemma to overcome in building trustworthy AI systems [10, 11, 9]: on the one hand, we expect AI systems to decide autonomously and intelligently on our behalf, which requires agency and delegation; on the other hand, we require them to be predictable, verifiable, safe and accountable. Of course, there are limits to achieving all
these goals and to guarantee correctness under all circumstances and domains. Instead, there is a trade-off to be made between entirely predictable and correct versus plausible and adaptive behaviour. What matters most is that expectations are managed to create validated trust through experience, shared causal models and theory of mind. Therefore, mutual awareness of assumptions, intentions, expectations and capabilities are required to create a dialogue of trust in human-agent collaboration.

A fundamental difference between traditional software and AI systems is that the outcomes are not necessarily deterministic, but probabilistic, and that there may be more than one "correct answer". Hence, the goal is shifting from guarantees of correctness towards verifying for plausibility. In the section below we discuss the five pillars of our trustworthy AI engineering framework.

2.1. Models and Explanations

Explicit models\(^2\) of the world or a suitable context in question enable reliable predictions of the behaviour of AI systems, both in the scope of training and outside, because they generalise knowledge beyond the limited and biased scope of the training data. Given a certain context, which can be very narrow or broad, explicit models represent concepts, relationships and rules that are always true in that context. For example, the laws of gravity are applicable to the whole universe. Models can be created by experts or learned from experience and data. Combinations of different types of models are particularly useful and insightful. For example, neuro-symbolic approaches are used to achieve this [13, 14, 15, 16, 17]. In this way, models promote transparency and explainability and, thus, make it possible to render the behaviour of the AI systems understandable and plausible. In simulations, models can enable the understanding – through experiments – of situations that are difficult or impossible to access otherwise. Often, synthetic data can be used to maintain privacy and avoid dangerous conditions.

Because models depend on a given context or domain, it is essential that agents using those models are aware of their competence in the given situation and are able to apply suitable models or adapt to situations gracefully when changing or leaving their scope of competence.

2.2. Causality and Grounding

Causality refers to the ability to identify and predict cause-and-effect relationships, i.e. which effects are the results of which causes and why [18]. An AI system that can understand causal relationships is able to make informed predictions and solve complex problems.

The need to move from correlation to causation is increasingly urgent (see figure 1, where the dotted line indicates correlation and the arrows indicate asymmetric causal relationships). If we want to explain why certain predictions are made or decisions are taken, it is essential to know and act on their causes.

---

\(^2\)The term 'model' is used extensively in the ML community. It is necessary, however, to distinguish between the statistical models of ML and the semantic models of knowledge engineering. Here, we refer to the latter. See also in [12] for a unified taxonomy of AI.
Causal inference is concerned with the resulting effect when a corresponding event (cause) occurs, according to a given causal model. Causal inference asks whether an event indeed causes a certain effect by determining the likelihood that one event was the cause of another. In contrast to statistical correlations, causal relationships are directed and asymmetrical.

Counterfactuals refer to alternative choices that could have been made in the past and the corresponding effects that they might have caused. Therefore, they allow for exploring possibilities to find alternative outcomes according to a causal model, allowing to change policies accordingly in the future.

Causal discovery allows for determining whether a change in one variable (representing a state, action or event) indeed causes a change in another, in order to distinguish between correlated and causal relationships in data and to derive corresponding causal models.

Closely related to causality is understanding the anchoring (grounding) of meanings in the real context. A deep understanding of context and meaning requires not only processing data, but also capturing the real-world phenomena that the data represents, such that predictions, decisions and actions are based on them. Layers of abstractions are fundamental for building rich architectures. Semantic models, such as ontologies, are representations of concepts, their attributes and relationships. They contribute to trustworthy AI systems by explaining and constraining the meaning of those concepts.

2.3. Modularity and Compositionality

One of the fundamental design principles of (software) engineering is modularity. Modularity guarantees that complex systems are broken down into understandable and manageable parts (functions and features) and assembled into system architectures. This increases the reliability of the individual components and their assemblies as systems of systems. It is much easier to verify smaller components than big monolithic artefacts. The evolution in software engineering from structured to modular and object-oriented programming enabled the design and construction of complex systems. In well-designed systems the transitions between successive components can be controlled and protected, making them explainable such that errors can be detected and repaired effectively. The pre- and post-conditions of each component can be validated and orchestrated in increasingly complex systems of systems.

When designing trustworthy AI systems, there are several important aspects that should be considered to guarantee the characteristics of trustworthy AI. In principle, these aspects apply to all software systems. However, they are of the greatest relevance for complex, intelligent systems for critical applications. AI engineering should make use of the lessons learned from software engineering and apply its engineering principles, such as design patterns and architectures.

An attempt to model design patterns for neuro-symbolic systems is made in [12, 19]. Two examples (see figure 2) show data-driven and knowledge-driven patterns. They are based on a visual language and taxonomy.

An important advantage of modular systems is that compositional patterns of subsystems can be identified and defined, which increases their reliability and documentation through reuse [20].
It is important to stress that software architectures are not merely static artefacts, but they rely on the interplay between structures and events – the organising principles and the dynamic evolution of complex systems [21]. Neither structure nor events are meaningful on their own, but require and depend on each other. In an extrapolated view, this relationship may be applied to the combination of learning and reasoning. Meaning emerges from a system’s structure and its components, when it is operated in a dynamic context of perceiving and acting.

The principle of compositionality also applies to knowledge models and languages [22]: larger constructs are created by joining together smaller units with specific, understandable, and verifiable tasks. Abstract relationships can thus be traced back to their components. These aspects are applied when designing complex systems and should also become a matter of course for AI systems.

2.4. Human Agency and Oversight

Human agency and oversight mean that, in any case, a human being should have the overview, final decision, and responsibility for the actions of an AI system (human empowerment). Even if many tasks are increasingly being transferred to autonomous AI systems (agents), the principle that humans supervise, assess, and approve actions still applies. Keeping in mind the above-mentioned dilemma in building trustworthy AI systems, delegation of tasks needs to be interpretable by both humans and (software) agents – in particular, when humans and agents collaborate as hybrid teams in a symbiotic partnership. It is necessary that suitable task descriptions are handed over to the agents and that they understand and execute them in the relevant context, considering the models, explanations and causal relationships explained above.

When considering the collaboration and competition in hybrid teams of humans and autonomous agents, we consider many-to-many situations where multiple humans and multiple agents form hybrid teams. The purpose of the agents is to empower humans with providing their complementary capabilities, such as fast and precise information exchange and analysis of large data sets. Agents can play many different roles, but the responsibility for decisions remains, in principle, with humans, for example by verifying, validating and approving proposals for decisions. An essential aspect of meaningful collaboration is to make mutual assumptions and expectations explicit, such that they can be used in deliberation and communication. This is a prerequisite for appropriate delegation of tasks and the accurate and concise descriptions of their underlying intentions.

For collaborative decision-making (CDM), it is essential that each human and agent is aware of each others’ points of view and has a notion of the others possess points of view that might differ from one’s own - which is known as a Theory of Mind (ToM). ToM is defined as the human cognitive ability to perceive and interpret others in terms
of their mental states and it is considered an indispensable requirement of human social life [23, 24, 25, 26, 27, 28]. Rather than reasoning only with one’s own beliefs, desires, goals, intentions, emotions, and thoughts, a person or agent with the awareness of others’ states of mind can consider different and mindful acts, depending on a perceived context. This ability allows them to more easily understand, predict, and even manipulate the behaviour of others [29].

Trustworthiness in interacting with artificially intelligent systems emerges from experience and as a combination of various properties, such as fairness, robustness, transparency, verification, and accuracy [30]. AI systems are trusted when we have confidence in the decisions that they take, i.e. when we understand why they are made [31], even when we disagree.

2.5. Maturity Models

Maturity of software is commonly denoted by means of Maturity Models3, which are frameworks that can be used in the process of planning and engineering, as well as in the process of road mapping. An extensive overview of AI maturity models is written by Sadiq et al [32].

In previous work [33] we presented a maturity-model that expressed the level of cooperation of an AI entity (agent) that acts in a human-AI team. The model can be used to reflect on expected capacities, roles and responsibilities of AI entities that act in a Human-AI team. The maturity model classifies the level of cooperation along two dimensions, level of agency and level of communication. It was represented by means of a two-dimensional matrix.

Horizontally, levels of agency range from human-controlled to fully autonomous [34, 35, 36]. Artificial Intelligence is based on the principles of autonomy and agency. Autonomy, the quality or state of being self-governing, is required to avoid purely predictable and reactive behaviour [37]. Whenever an AI entity commits on contributing to a team intention, it has to balance its level of autonomy with required levels of interaction. The model distinguishes four levels of agency with respect to an AI entity acting in a team (Human Trusted, Situational Autonomy, Preferred Autonomy and AI Trusted).

Vertically, levels of communication vary from merely sharing information about alignment of tasks, to forms of interaction in which an agent takes into account the mental state of others. Communication is a prerequisite for deliberation, delegation of tasks and sharing of knowledge within an ecosystem. We distinguish four levels of communication, varying from simple sharing information about coordination of tasks to higher orders of interaction that include the exchange of information about the learning process and each others’ mental states. The maturity model distinguishes four levels of communication (Task Alignment, Co-Learning, Mental Modelling and Motivating).

We extended the maturity model with a third axis (see figure 3). The maturity model, named the AI Interaction Maturity Model, manifests itself in the form of a three-dimensional cube. the third axis can be used to express the impact of a particular AI team member on the social level, varying from an individual level (silo-ed situation) to organisational, industrial domain-specific and societal level.

---

3 Another note on the term ‘model’. as it is an ambiguous term; where-as in software engineering it may refer to a computational model, theoretical model or architectural model, we use the term model here to refer to an engineering framework, commonly denoted with the term ‘maturity model’.
Figure 3. The AI Interaction Maturity Model extends the Collaborative Agent Maturity Model by including a third axis that reflects the impact of an AI entity on the social level.

Figure 4. A causal model for living and working in the urban context.

3. Trusted Urban AI Scenario

As an example for applying the above-mentioned AI engineering concepts, we propose a scenario in an urban context.

Urban life has many peculiar characteristics [38, 39, 40, 41, 42, 43]. Some causal relationships in an urban context, focusing on energy consumption and mobility, are shown in figure 4. We are concerned with modelling and understanding human behaviour in an urban context (sustainable smart city), particularly when humans are part of hybrid teams together with agents.
The interactions in an urban environment are diverse, complex and conflicting. Many interests of hybrid actors are related and depend on each other. In the urban context, an overall goal for sustainable use of resources could be the reduction of energy consumption.

Using causal models, adopting and implementing the engineering disciplines and using a maturity model, as explained above, improves our understanding and control of systems that we design and apply in this urban context. The consumption of various types of energy is affected by the need and desire to move about the city and to heat buildings at home and at work (and for leisure, shopping, etc.). As shown in fig. 4, values and duties are the main sources that drive urban behaviour and external factors, such as the weather, influence decision-making. This causal model explains the relationships among several important behavioural aspects, but it is not deterministic. Individual behaviour is influenced by exogenous variables and cooperative behaviour results in complex interactions. A shared goal can be seen and modelled as an effect, that is caused by one or more interventions (actions or events). Consequently, in order to decide and plan which actions to take, it is necessary to understand which actions or events cause the intended effects. For example, your goal can be to arrive at a destination at a given time (work, home, leisure, etc.). By reasoning back which actions are required to get you there, piece by piece, a connected causal path can be constructed to determine the departure time and modes of traffic along the route. Due to shared intentions and causal models, humans and agents can mutually trust each other regarding their actions and outcomes.

The urban context comes with a multidisciplinary stakeholder field, involving a landscape of IT systems varying from traditional components in data sharing platforms to AI-enabled services. A modular approach for both design and realisation of software modules and AI models is crucial to keep the systems at their required levels of interoperability and scalability. Roadmapping of AI based systems, inside buildings as well as between various buildings and their interaction with human engineers is facilitated by means of the AIMM model.

4. Conclusions

As the field of Artificial Intelligence is still, and again, facing tremendous and overwhelming changes and progress, there is a strong and quickly growing need for trust in AI systems. The goal of Trust by Design is proposed to be based on the five engineering principles of (i) Models & Explanations, (ii) Causality & Grounding, (iii) Modularity & Compositionality, (iv) Human Agency & Oversight, and (v) a Maturity Model. Our intention is to develop the insights above further into practical methods and tools, based on a design pattern language, to benefit the AI community and its users. The context of energy consumption and mobility in an urban context serves as applied setting setting in various projects, validation of our suggested 5 pillars for controlled engineering and further experimentation in the field of human-AI ecosystems.

Acknowledgements

This research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.
References


33. Mulder W and Meyer-Vitali A. A Maturity Model for Collaborative Agents in Human-AI Ecosystems. *Collaborative Networks in Digitalization and Society 5.0*. Ed. by Camarinha-Matos LM, Boucher X, and Ortiz A. IFIP Advances in Information and Communication Technology. Cham: Springer Nature Switzerland, 2023:328–35. DOI: 10.1007/978-3-031-42622-3_23


Hybrid Intelligence in Academic Writing: Examining Self-Regulated Learning Patterns in an AI-Assisted Writing Task

Andy NGUYEN, Faith ILESANMI, Belle DANG, Eija VUORENMAA, and Sanna JÄRVELÄ
Learning and Educational Technology (LET) Lab, University of Oulu, Finland
ORCiD ID: Andy NGUYEN https://orcid.org/0000-0002-0759-9656

Abstract. The arrival of generative Artificial Intelligence (AI) in educational settings offers a unique opportunity to explore the intersection of human cognitive processes and AI, especially in complex tasks like writing. This study adopts a process-oriented approach to investigate the self-regulated learning (SRL) strategies employed by 21 doctoral and master's students during a writing task facilitated by generative AI. It aims to identify and analyze the SRL strategies that emerge within the framework of hybrid intelligence, emphasizing the collaboration between human intellect and artificial capabilities. Utilizing a learning analytics methodology, specifically lag sequential analysis (LSA), the research examines process data to reveal the patterns of learners' interactions with generative AI in writing, shedding light on how learners navigate different SRL strategies. This analysis facilitates an understanding of how learners adaptively manage their writing task with the support of AI tool. By delineating the SRL strategies in AI-assisted writing, this research provides valuable implications for the design of educational technologies and the development of pedagogical interventions aimed at fostering successful human-AI collaboration in various learning environments.

Keywords. Hybrid Intelligence, AI in education, Human-AI Interaction, Academic Writing, Higher Education, AIED

1. Introduction

The integration of Artificial Intelligence (AI) into educational contexts has marked a significant shift in the landscape of learning and instruction, particularly with the emergence of generative AI technologies. These advancements offer profound opportunities for enhancing educational experiences, especially in the development of complex cognitive skills such as writing. Writing, an essential academic skill, involves a myriad of cognitive processes, including planning, drafting, revising, and editing. The advent of generative AI in educational settings presents a novel avenue for exploring the synergy between human cognitive capabilities and artificial intelligence, especially in the realm of writing tasks. This exploration is crucial, as writing not only is a key academic skill but also serves as a window into students' cognitive processes, offering insights into how they organize, express, and refine their thoughts [1,2].

1 Corresponding Author: Andy Nguyen, Andy.Nguyen@oulu.fi.
Self-regulated learning (SRL) strategies play a pivotal role in successful writing, as they enable learners to manage their cognitive, metacognitive, and affective processes effectively [3–5]. SRL strategies are particularly relevant in tasks that require high levels of cognitive engagement and autonomy, such as writing. The integration of generative AI tools in writing tasks introduces a new dimension to SRL, as learners must navigate the interplay between their cognitive strategies and the capabilities of AI technologies. This intersection forms a hybrid intelligence system, where human cognitive processes and artificial intelligence collaborate to accomplish tasks [6,7].

Despite the potential of generative AI to transform educational practices, empirical research exploring how learners adapt their SRL strategies within AI-assisted environments, particularly in complex tasks like writing, remains limited. Understanding how learners employ SRL strategies in conjunction with generative AI tools can provide valuable insights into the nature of human-AI collaboration in different learning contexts. Moreover, examining the impact of these patterns is essential for design the effective hybrid intelligence systems for writing tasks.

This study aims to fill this gap by adopting a process-oriented approach to investigate the SRL strategies employed by doctoral and master's students during a writing task involving the use of generative AI tool(s). By focusing on the process data through learning analytics methodologies, specifically sequential clustering and lag sequential analysis (LSA), this research seeks to uncover the patterns of learners' interactions with generative AI tools. In particular, this study aims to address the following research question: What are the key patterns of actions, embedded with self-regulated learning (SRL) features, in an AI-assisted writing task?

The significance of this study lies in its potential to advance understanding of the effective integration of AI in educational practices such as academic writing. By identifying and analyzing the SRL strategies that emerge in the context of hybrid intelligence, this research contributes to the development of pedagogical interventions and the design of educational technologies that support successful human-AI collaboration. Furthermore, the findings of this study offer implications for educators and technology designers in fostering environments that leverage the strengths of both human intellect and artificial capabilities to enhance learning outcomes, particularly in the domain of academic writing.

2. Theoretical Foundations

This section outlines the theoretical framework guiding our study, focusing on three key areas: 1) SRL theory and SRL research in the context of academic writing, 2) related work on intelligent writing support systems, and 3) the adopted human-AI shared regulation (HASRL) framework for hybrid intelligence.

2.1. Self-regulated Learning Process Features in Academic Writing Tasks

The concept of self-regulated learning (SRL) has emerged as a pivotal element in the educational research landscape, signifying the capacity of learners to autonomously navigate and control their learning endeavors. At the macro-level, this self-regulatory ability encompasses a range of activities, including setting learning goals, employing strategies to achieve these goals, and monitoring progress towards them [3]. The COPES model, as articulated by Winne and Hadwin [4], presents a more detailed framework for
understanding SRL through a structured analysis of its components. This model delineates SRL into five interrelated elements: Conditions, Operations, Products, Evaluations, and Standards, collectively encompassing the multifaceted nature of self-regulation in learning contexts. The significance of SRL in enhancing educational outcomes has been well-documented across various stages of learning and diverse educational contexts. The literature suggests that learners who adeptly manage their learning processes tend to achieve higher academic success and display greater persistence in the face of challenges [8,9].

Recent advancements in learning analytics and advanced technologies have further enriched our understanding of SRL by providing nuanced insights into the micro-level processes of self-regulation. These technological innovations enable the examination of learner behaviors and strategies at a granular level, thereby offering a more detailed picture of the self-regulatory processes in action [7,10]. Particularly, multimodal learning analytics have been instrumental in dissecting the complex nature of SRL, allowing researchers to explore how learners interact with various educational tasks and processes through different modalities [11,12].

In the area of academic writing, a critical skill in higher education, the exploration of SRL process features has gained momentum. Recent studies have begun to unravel how learners regulate their writing processes [13–15], employing strategies such as planning, monitoring, and revising to enhance their written outputs. For instance, Rakovic et al. [16] have demonstrated the potential of utilizing trace and process data and linguistic analysis to predict learner performance in multi-text writing tasks. This approach highlights the value of process data, including digital trace data, in uncovering the metacognitive aspects of writing, offering insights into how learners source, integrate, and synthesize information from multiple texts.

Moreover, the integration of generative AI and advanced learning analytics into the writing process presents both opportunities and challenges for SRL. The arrival of these technologies has sparked a discourse on their implications for traditional writing practices, highlighting the need for further investigation into how human-AI collaboration can be leveraged to foster hybrid intelligence in learning. Such a model aims to optimize learning outcomes by combining the strengths of human cognitive processes with the computational power of AI [7].

However, the rapid evolution of generative AI tools in educational settings has also raised concerns regarding their potential to disrupt established learning and writing processes. These concerns necessitate a deeper examination of the ways in which AI can be integrated into the learning process without undermining the essential components of SRL, such as goal setting, strategy use, and self-monitoring. As we move forward, it is essential to explore how learners should best perform learning with the technologies [17].

2.2. Intelligent Writing Support Systems

The development of intelligent writing support systems, aimed at enhancing learners' writing skills, is not a new concept, with efforts spanning over the last decade to both design these technologies and research their impact on writing processes. The evolution of these systems has been marked by significant advancements, ranging from grammar checking tools, such as Grammarly [18], to more sophisticated platforms designed to support specific aspects of the writing process. For instance, G-Asks represents a notable development, offering an intelligent automatic question generation system tailored for academic writing support [19]. This diversity in tools reflects a broad interest in
leveraging technology to facilitate the complex task of writing, underscoring the potential of intelligent writing support systems to act as pivotal aids in the educational landscape.

The advent of advanced generative AI technologies has introduced a new dimension to the discourse on intelligent writing support systems, particularly concerning their application in academic writing. Generative AI’s ability to produce human-like text has sparked both enthusiasm and concern regarding its implications for writing processes and outcomes. This technological leap forward raises critical questions about the integration of generative AI within intelligent writing support systems and its potential to transform writing practices in educational settings. Recent discussions have engaged various stakeholders, including teachers, learners, and educational experts, to explore perspectives on the utilization of generative AI in academic writing [20,21]. These conversations have highlighted a spectrum of views, from optimistic appraisals of AI’s supportive role to cautious evaluations of its impact on learners’ writing skills.

Despite the growing interest in the capabilities of generative AI and its integration into intelligent writing support systems, there remains a notable gap in empirical research, particularly concerning process-oriented investigations of writing processes with generative AI. Such research is essential for understanding not only the outcomes of using generative AI in writing tasks but also the dynamics of interaction between human learners and AI systems. A critical area of inquiry involves examining how learners can effectively collaborate with generative AI to optimize writing outputs, ensuring that the use of AI enhances rather than diminishes their writing skills. The challenge lies in identifying strategies that leverage the strengths of generative AI while fostering the development of learners’ own abilities [7,22], a balance crucial for maintaining the educational value of writing tasks.

While intelligent writing support systems, including those powered by generative AI, offer promising avenues for supporting academic writing, the field stands at a crossroads. The potential of these technologies to revolutionize writing practices necessitates a deeper, process-oriented exploration of how they can be integrated into learning environments to benefit rather than hinder the development of writing skills. As the landscape of educational technology continues to evolve, further research is imperative to guide the effective and ethical use of generative AI in academic writing.

2.3. Towards Human-AI Shared Regulation for Hybrid Intelligence

The context of increasing AI adoption has led to the emergence of the concept of Hybrid Intelligence as a promising approach for integrating human and machine capabilities [6]. Hybrid Intelligence envisages a collaborative model where humans and AI systems work together, leveraging their respective strengths to achieve superior outcomes than either could accomplish alone. In the realm of education, particularly in self-regulated learning (SRL), the potential for Hybrid Intelligence through human-AI collaboration has been recognized by leading scholars [7,22]. Järvelä et al. [7] introduced the Human-AI Shared Regulation (HASRL) framework as a means to explore the self-regulatory processes within the context of human-AI collaboration, aiming to foster a synergistic relationship between human cognitive abilities and artificial intelligence.

This study adopts the COPES and HASRL framework as its theoretical foundation, focusing specifically on the traces of human SRL process features within HASRL model. While the design and development of AI systems within the HASRL framework are beyond the scope of this research, the investigation into human SRL processes in
academic writing, facilitated by generative AI, provides critical insights. These insights are not only valuable in understanding the dynamics of human SRL in conjunction with AI but also hold the potential to guide the future design and development of Hybrid Intelligence systems. By examining the interplay between human SRL process features and AI in an academic writing task, this study contributes to the broader discourse on the optimal integration of AI in educational contexts, aiming to enhance learning outcomes through the strategic combination of human and machine intelligence.

3. Methods

This study employs a process-oriented learning analytics approach to investigate the process data from learners using generative AI for a writing task. This method allows for the examination of how learners dynamically interact with AI tools during the writing process, aiming to uncover insights into the SRL process features in writing with generative AI. This approach is similar to prior studies in SRL research with learning analytics [10,23,24].

3.1. Participants and Context

The participants in this study consisted of 21 graduate students, including seven PhD candidates and fourteen master's students, all enrolled in Learning and Educational Technology programs. This group was selected for their extensive experience with academic writing and the significant role that such writing plays in their academic and professional development. Insights derived from this cohort are anticipated to illuminate the dynamics of human-AI collaboration in sophisticated academic writing, which requires high standards and the articulation of complex ideas.

The data collection of this investigation was designed as an online observational study conducted via the Zoom video conferencing platform, reflecting the prevalent trends in digital communication and remote collaboration. Participants were involved in a 30-minute writing task, which required composing a short essay of approximately 500 words on the use of AI in education. This task demanded that participants express their viewpoints, supported by evidence and examples, akin to the expectations of academic writing assignments encountered in their studies.

To ensure the authenticity of the writing experience and its relevance to their academic pursuits, the design of the writing task and the evaluation rubric were closely aligned with those typically used in university settings. Participants were given a 30-minute period to complete the essay, during which they were allowed to use any resources they deemed necessary, including ChatGPT, Google Bard, Quillbot and Google Scholar. This setup aimed to replicate real-world writing scenarios, facilitating an examination of the participants’ disposition to employ AI in writing tasks.

Experimental sessions were recorded to document participants' interactions with the writing task and tool usage, offering insights into their behaviors and chosen resources. A pre-survey questionnaire gathered participants' background information before each session. Identifying details were anonymized in recordings and responses to protect privacy. Informed consent was secured from all participants, ensuring the study adhered to university ethics and GDPR standards.
3.2. Qualitative Coding Analysis

Screen recordings data were analyzed and qualitatively coded to identify patterns in students' use of AI-assisted writing tools. Employing qualitative content analysis with the constant comparison method [25], each student's writing-related action was initially tagged with descriptive codes in the open coding phase. These codes were then grouped into broader categories, leading to the formulation of themes that captured the essence of the writing behaviors observed. This approach allowed for a detailed examination of the micro-processes students engaged in while writing, drawing on previous studies that explored SRL process features [23,26], specifically in academic writing tasks [15]. Table 1 shows the coding scheme for the actions recorded in this study.

<table>
<thead>
<tr>
<th>Action</th>
<th>Code</th>
<th>Action description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Review Instruction</td>
<td>INSTRUCT</td>
<td>Read, re-read, or review general instructions, task requirements, and the rubric.</td>
</tr>
<tr>
<td>Search Information</td>
<td>SEARCH</td>
<td>Conduct searches for words, concepts, and articles using non-generative AI tools, such as Google, Google Scholar, or other browsers.</td>
</tr>
<tr>
<td>Prompt GenAI</td>
<td>PROMPT</td>
<td>Engage generative AI platforms like ChatGPT, Google Bard, Bing Chat and prompt for information, content, feedback, references, etc.</td>
</tr>
<tr>
<td>Review GenAI Content</td>
<td>RE_GAI</td>
<td>Read, re-read or review information and content generated by generative AI.</td>
</tr>
<tr>
<td>Read Article</td>
<td>ARTICLE</td>
<td>Read, re-read or review article content.</td>
</tr>
<tr>
<td>Copy Paste Content</td>
<td>PASTE</td>
<td>Copy and paste content from non-generative AI sources (articles, notes, web pages, etc.) directly into the essay.</td>
</tr>
<tr>
<td>Copy Paste GenAI Content</td>
<td>PASTE_GAI</td>
<td>Copy and paste generated content or references from generative AI (ChatGPT, Google Bard, Bing Chat) directly into the essay.</td>
</tr>
<tr>
<td>Write Essay</td>
<td>WRITE</td>
<td>Write, edit, format, or stay in the essay zone for reviewing essay content.</td>
</tr>
<tr>
<td>Check Word Count</td>
<td>COUNT</td>
<td>Check the current word count of the essay.</td>
</tr>
<tr>
<td>Reference</td>
<td>REF</td>
<td>Manually or through applications such as Google Scholar, Mendeley, Zotero, etc., incorporate scholarly citations and references into the essay.</td>
</tr>
</tbody>
</table>

3.3. Sequential Clustering

To analyze the main patterns of AI-assisted writing, this study applied Agglomerative Hierarchical Clustering (AHC) using an optimal matching algorithm (OM) [27]. The analysis was conducted in Python, utilizing the scikit-learn library for its advanced machine learning functionalities. AHC, an unsupervised learning method, aims to uncover the dataset's inherent structure by grouping data points into clusters based on similarity. The optimal number of clusters was determined through the analysis of Silhouette Coefficient and a dendrogram. This approach enabled the identification of
distinct writing behaviors among participants, providing valuable insights into the diverse strategies utilized in AI-assisted writing tasks.

3.4. Lag Sequential Analysis

To investigate the sequential dynamics of activities embedding SRL process features within clusters, lag sequential analysis (LSA) was utilized. This method allowed for the calculation of transitional probabilities through overlapped sampling, assessing how likely it was for certain activities to follow one another. LSA aimed to pinpoint event sequences occurring more frequently than would be expected by chance, highlighting meaningful patterns in SRL behavior. The term "lag" in LSA refers to the position of an event relative to another, with "lag 1" indicating direct sequences where events immediately follow each other, and "lag 2" highlighting indirect sequences with an intermediary action, suggesting complex or delayed learning strategies. The analysis employs a likelihood ratio chi-square ($\chi^2$) to determine if sequence frequencies deviate significantly from expected values. When the overall $\chi^2$ for an analysis is significant, each observed sequence can be evaluated through adjusted residuals and $z$-scores. Sequences with the acceptable $z$-scores $\geq \pm 1.96$ ($P = 0.05$) and Yule's $Q$ value of at least 0.30 for association strength are considered significant [28]. This analysis sheds light on the structured progression of learning behaviors, providing insight into the temporal understanding of how learners navigate in AI-assisted writing.

4. Results and Findings

The outcomes of the sequential clustering analysis delineate the categorization of graduate students' behavioral patterns in AI-assisted writing tasks into two clusters. The descriptive statistical attributes of these clusters are systematically presented in Table 2.

Table 2. Descriptive statistics for two clusters detected with Agglomerative Hierarchical Clustering (AHC)

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sequences</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>Average sequence length</td>
<td>81.89 (SD 13.45)</td>
<td>47.83 (SD 15.00)</td>
</tr>
<tr>
<td>Total actions</td>
<td>737</td>
<td>574</td>
</tr>
<tr>
<td>1 - Review Instruction</td>
<td>38</td>
<td>39</td>
</tr>
<tr>
<td>2 - Search information</td>
<td>89</td>
<td>63</td>
</tr>
<tr>
<td>3 - Prompt GenAI</td>
<td>69</td>
<td>62</td>
</tr>
<tr>
<td>4 - Review GenAI Content</td>
<td>138</td>
<td>77</td>
</tr>
<tr>
<td>5 - Read Article</td>
<td>102</td>
<td>74</td>
</tr>
<tr>
<td>6 - Copy Paste Content</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>7 - Copy Paste GenAI</td>
<td>54</td>
<td>40</td>
</tr>
<tr>
<td>8 - Write Essay</td>
<td>149</td>
<td>144</td>
</tr>
<tr>
<td>9 - Check Word Count</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>10- Reference</td>
<td>87</td>
<td>51</td>
</tr>
</tbody>
</table>

Cluster 1, characterized by a higher total number of actions in each sequence, indicates a more intensive interaction with the writing task, as evidenced by a longer average sequence length and a greater total number of actions compared to Cluster 2. Specifically, Cluster 1 shows a more frequent engagement with activities such as reviewing AI-generated content and reading articles, suggesting a deeper research and...
review process. Conversely, Cluster 2 exhibits shorter sequence lengths on average and a lower total action count, pointing towards a more concise and perhaps more efficient approach to the writing task. Notably, Cluster 2 participants were more likely to review instructions and check word count which may indicate task understanding and monitoring behavior, an essential SRL feature [15,29]. These differences highlight the variability in how graduate students utilize AI tools in academic writing, with some favoring extensive research and content generation, while others prioritize editing and refining their drafts.

For Cluster 1, Table 3, detailing transitional probabilities, Figure 1 illustrates the likelihood of transitioning between different activities during the AI-assisted writing process. The chi-square analysis confirms a significant association between the rows and columns within the frequency counts ($\chi^2 = 670.15$, degrees of freedom = 81, $p < .001$).

**Table 3. Transitional Probabilities (Cluster 1)**

<table>
<thead>
<tr>
<th>Activity</th>
<th>INSTRUCT</th>
<th>PROMP T</th>
<th>RE_GA I</th>
<th>PASTE_GA I</th>
<th>WRITE</th>
<th>SEARCH</th>
<th>ARTIC LE</th>
<th>REF</th>
<th>PASTE</th>
<th>COUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSTRUCT</td>
<td>2.63</td>
<td>31.58</td>
<td>23.68</td>
<td>0.00</td>
<td>23.68</td>
<td>5.26</td>
<td>5.26</td>
<td>7.89</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>PROMPT</td>
<td>7.25</td>
<td>0.00</td>
<td>62.32</td>
<td>11.59</td>
<td>5.80</td>
<td>4.35</td>
<td>7.25</td>
<td>1.45</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>RE_GAI</td>
<td>6.52</td>
<td>18.12</td>
<td>2.17</td>
<td>25.36</td>
<td>28.99</td>
<td>8.70</td>
<td>9.42</td>
<td>0.72</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>PASTE_GAI</td>
<td>1.85</td>
<td>11.11</td>
<td>29.63</td>
<td>0.00</td>
<td>38.89</td>
<td>7.41</td>
<td>1.85</td>
<td>1.85</td>
<td>0.00</td>
<td>7.41</td>
</tr>
<tr>
<td>WRITE</td>
<td>8.05</td>
<td>8.05</td>
<td>29.53</td>
<td>6.04</td>
<td>1.34</td>
<td>5.37</td>
<td>16.11</td>
<td>22.15</td>
<td>2.01</td>
<td>1.34</td>
</tr>
<tr>
<td>SEARCH</td>
<td>6.74</td>
<td>4.49</td>
<td>4.49</td>
<td>0.00</td>
<td>10.11</td>
<td>0.00</td>
<td>50.56</td>
<td>23.60</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ARTICLE</td>
<td>1.96</td>
<td>7.84</td>
<td>14.71</td>
<td>1.96</td>
<td>19.61</td>
<td>26.47</td>
<td>0.98</td>
<td>25.40</td>
<td>0.98</td>
<td>0.00</td>
</tr>
<tr>
<td>REF</td>
<td>1.15</td>
<td>1.15</td>
<td>1.15</td>
<td>0.00</td>
<td>45.98</td>
<td>36.78</td>
<td>12.64</td>
<td>0.00</td>
<td>1.15</td>
<td>0.00</td>
</tr>
<tr>
<td>PASTE</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>50.00</td>
<td>25.00</td>
<td>0.00</td>
<td>25.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>COUNT</td>
<td>0.00</td>
<td>16.67</td>
<td>50.00</td>
<td>0.00</td>
<td>33.33</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>0%</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.** Cluster 1 state transition diagram of signification lag 1 transition ($z > 1.96, Q < 0.30$). Edges are labeled with z-score. Lag 1 is continuous line; Lag 2 is dash line.

A notable pattern emerges from Cluster 1, indicating a high probability of moving forward and backward from "PROMPT" to "RE_GAI" (62.32%, $z = 9.74$, $Q = 0.82$), suggesting that after prompting generative AI, students are most likely to review the AI-generated content. This transition underscores the significant role of AI in shaping the initial stages of the writing process. The "WRITE" activity also shows a diverse range of
subsequent activities, with the highest probabilities indicating indirect transitions back to "WRITE" (29.53%, $z_{lag2} = 7.85$, $Q_{lag2} = 0.64$) and direct transition to "REF" (22.15%, $z = 4.38$, $Q = 0.47$), suggesting iterative processes of writing and referencing. Interestingly, the transition from "PASTE_GAI" to "WRITE" ($z = 3.55$, $Q = 0.47$) is relatively high (38.89%), pointing to a significant reliance on integrating AI-generated content into the writing process. Conversely, activities like "PASTE" and "COUNT" show limited transitions, indicating these actions are less central to the workflow. These findings highlight the pivotal role of AI in facilitating the writing process, particularly in the generation and refinement of content, while also pointing to an iterative cycle of writing, referencing, and revising as key components of the students' writing strategies.

For Cluster 2, Table 4, detailing transitional probabilities, Figure 2 illustrates the likelihood of transitioning between different activities during the AI-assisted writing process. The chi-square test reveals a significant correlation between the rows and columns in the frequency data ($\chi^2 = 498.15$, df = 83, $p < .001$).

### Table 4. Transitional Probabilities (Cluster 2)

<table>
<thead>
<tr>
<th>Activity</th>
<th>INSTRUCT</th>
<th>PROMPT</th>
<th>RE_GAI</th>
<th>WRITE</th>
<th>PASTE_GAI</th>
<th>SEARCH</th>
<th>ARTICLE</th>
<th>REF</th>
<th>PASTE</th>
<th>COUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>INSTRUCT</td>
<td>10.26</td>
<td>38.46</td>
<td>12.82</td>
<td>23.08</td>
<td>2.56</td>
<td>5.13</td>
<td>2.56</td>
<td>2.56</td>
<td>2.56</td>
<td>0.00</td>
</tr>
<tr>
<td>PROMPT</td>
<td>12.90</td>
<td>0.00</td>
<td>37.10</td>
<td>11.29</td>
<td>24.19</td>
<td>6.45</td>
<td>4.84</td>
<td>1.61</td>
<td>1.61</td>
<td>0.00</td>
</tr>
<tr>
<td>RE_GAI</td>
<td>7.79</td>
<td>12.99</td>
<td>0.00</td>
<td>38.96</td>
<td>19.48</td>
<td>10.39</td>
<td>6.49</td>
<td>3.90</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>WRITE</td>
<td>11.89</td>
<td>17.48</td>
<td>18.18</td>
<td>3.50</td>
<td>3.50</td>
<td>7.69</td>
<td>16.78</td>
<td>11.89</td>
<td>3.50</td>
<td>5.59</td>
</tr>
<tr>
<td>PASTE_GAI</td>
<td>0.00</td>
<td>10.00</td>
<td>12.50</td>
<td>57.50</td>
<td>0.00</td>
<td>12.50</td>
<td>2.50</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SEARCH</td>
<td>1.59</td>
<td>6.35</td>
<td>11.11</td>
<td>7.94</td>
<td>0.00</td>
<td>0.00</td>
<td>46.03</td>
<td>26.98</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ARTICLE</td>
<td>0.00</td>
<td>1.35</td>
<td>9.46</td>
<td>36.49</td>
<td>0.00</td>
<td>28.38</td>
<td>0.00</td>
<td>16.22</td>
<td>4.05</td>
<td>0.00</td>
</tr>
<tr>
<td>REF</td>
<td>0.00</td>
<td>5.88</td>
<td>5.88</td>
<td>45.10</td>
<td>0.00</td>
<td>21.57</td>
<td>19.61</td>
<td>0.00</td>
<td>0.00</td>
<td>1.96</td>
</tr>
<tr>
<td>PASTE</td>
<td>0.00</td>
<td>0.00</td>
<td>10.00</td>
<td>40.00</td>
<td>0.00</td>
<td>10.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>30.00</td>
</tr>
<tr>
<td>COUNT</td>
<td>14.29</td>
<td>0.00</td>
<td>0.00</td>
<td>78.57</td>
<td>7.14</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

In this cluster, a significant transition from "PROMPT" to "RE_GAI" (37.10%, $z = 5.79$, $Q = 0.67$) is also observed, albeit lower than in Cluster 1. However, the transition from "RE_GAI" to "WRITE" (38.96%, $z = 3.02$, $Q = 0.36$) is notably higher than in Cluster 1, suggesting a more direct path from reviewing AI content to writing. Similar to Cluster 1, the transition from "PASTE_GAI" to "WRITE" (57.50%, $z = 4.90$, $Q = 0.64$)
is also high, underscoring a stronger reliance on incorporating AI-generated content directly into the writing. Nevertheless, the "WRITE" activity in Cluster 2 shows a broader distribution of subsequent activities, with a significant indirect transition back to "WRITE" (3.50%, $z_{lag2} = 7.06$, $Q_{lag2} = 0.62$) and a prominent direct transition to "COUNT" (5.59%, $z = 5.69$, $Q = 0.61$), indicating a unique focus on monitoring word count during the writing process. Furthermore, direct transition "WRITE" back to "INSTRUCT" ($z = 2.67$, $Q = 0.45$) within this cluster, unseen in cluster 1, indicates more frequent monitoring of the task requirements. This cluster also exhibits a unique pattern in the "COUNT" activity, with a substantial indirect transition to "PROMPT" ($z_{lag2} = 2.16$, $Q_{lag2} = 0.56$) and direct transition to "WRITE" (78.57%, $z = 4.67$, $Q = 0.84$). This highlights a pattern of focusing on continuous monitoring of their writing progress and the then engagement with AI-assistance can be inferred as in response to this need.

Comparing the transitional probabilities between Cluster 1 and Cluster 2 reveals nuanced differences in how each cluster engages with AI-assisted writing tools. While both clusters demonstrate a reliance on AI for generating and refining content, Cluster 2 participants show a more streamlined approach from reviewing AI-generated content to writing. This is further emphasized by their higher propensity to transition from "PASTE_GAI" directly to "WRITE" and their unique attention to word count and reviewing instruction as part of their writing process. In contrast, Cluster 1 exhibits a more iterative process involving adopting AI-generated content, referencing, and revising, with a diverse range of activities following the writing process. These differences highlight distinct strategies employed by each cluster, with Cluster 2 leaning towards a more efficient, perhaps a more goal-oriented approach, while Cluster 1 engages in a more exploratory and iterative method.

Our results align with recent studies that investigate regulatory process features at the micro-level in the context of collaborative learning [15,23,26]. For instance, Dang et al. [30] identified two predominant regulatory strategies in collaborative learning: trial-and-failure and planning-and-implementation. Moving into the domain of individual writing tasks supported by generative AI, our study adds new insights into the learning process and reveals distinct SRL behaviors within this learning activity. This contribution deepens the understanding of SRL, showing how individuals engage with and manage their writing tasks when utilizing AI tools, thereby broadening the discussion on learning processes in the context of technological advancements.

5. Discussions

This research contributes to the understanding of how graduate students utilize AI in the context of academic writing, offering a glimpse into the future of educational technology where human-AI collaboration may become a cornerstone of learning processes [7,31]. The integration of AI in educational settings, as demonstrated by this study, holds the potential to significantly augment the SRL capabilities of learners, provided that the tools are used judiciously and in ways that complement human cognitive functions. The exploration of SRL process features within the context of human-AI collaboration, particularly through the case of graduate students engaging in the academic writing task.

This study contributes to the ongoing discussions within the learning sciences regarding the collaboration between humans and AI for shared regulation in learning [7,22]. This investigation also align with the broader discourse on the role of AI in education, suggesting that while advanced AI tools, such as ChatGPT and Google Bard,
offer substantial support for academic writing, the essence of human oversight and critical engagement remains irreplaceable [6,20,32,33]. The process-oriented learning analytics approach [23,30] facilitated a granular analysis of learner interactions with generative AI.

Our study shows the nuanced differences observed between the two clusters in their engagement with AI tools reflecting a broader spectrum of SRL characteristics that learners employ when interacting with technology. Specifically, the streamlined approach of Cluster 2, characterized by a direct transition from reviewing AI-generated content to writing, alongside a focused attention to word count, suggests a more efficiency-driven and goal-oriented strategy. Checking word count and task requirements aligns with the COPES model [4], showcasing students' SRL behavior [15]. This action reflects their strategic approach to meeting learning objectives by monitoring progress and adjusting efforts to align with set standards and expectations. In contrast, the iterative process observed in Cluster 1, involving extensive writing, referencing, and revising activities, indicates a preference for a more exploratory and iterative approach to learning and writing. Our findings raised a question regarding the effective integration of AI in educational practices necessitates a balanced synergy between technological affordances and human cognitive and metacognitive strategies [10,17].

The concept of Hybrid Intelligence, as evidenced through the Human-AI Shared Regulation (HASRL) framework proposed by Järvelä et al. [7], emerges as a promising paradigm for future educational technologies. This study's adherence to the HASRL framework, while focusing predominantly on the human aspects of SRL, illuminates the different SRL patterns of learners in complex cognitive tasks like academic writing with AI assistance. This study highlights the necessity of equipping learners with essential SRL skills to effectively collaborate with AI, rather than merely depending on it.

The implications of these findings are manifold. Firstly, they underscore the importance of designing AI tools that are flexible and adaptable to accommodate a wide range of learning strategies and preferences. Understanding the specific ways in which learners engage with AI to support their writing tasks can inform the development of more intuitive and supportive AI-based educational technologies. Secondly, this study highlights the potential of AI to serve as a catalyst for either enhancing or diminishing SRL. Educators can leverage these insights to guide the integration of AI tools into the curriculum, fostering environments that encourage effective SRL behaviors. Furthermore, the distinct patterns of AI tool usage identified in this study suggest the need for educational interventions that are tailored to different learning strategies. By recognizing the diversity in learner engagement with AI-assisted writing, educators and instructional designers can create more personalized learning experiences that enhance student motivation, engagement, and ultimately, learning outcomes.

6. Conclusions, Limitations, and Future Directions

In conclusion, this study not only illuminates the present state of collaboration between human intelligence and AI in academic writing but also invites further investigation into the transformative potential of AI in reshaping educational practices. By examining the nuanced ways in which students engage with AI to support their writing processes, this research contributes valuable insights into the evolving relationship between learners and AI for hybrid intelligence. It underscores the possibility of a synergistic partnership where AI tools not only assist in the mechanical
aspects of writing but also stimulate critical thinking, creativity, and deeper engagement with content. Such insights are crucial for understanding the current capabilities and limitations of AI in educational settings and for envisioning future directions where AI could play a pivotal role in fostering a more interactive and engaging learning environment. This study, therefore, not only highlights the state of human-AI collaboration but also emphasizes the need for ongoing research to fully realize the potential of AI in revolutionizing educational practices, making learning more accessible, engaging, and effective for students across diverse learning environments.

This study's limitations are highlighted by its relatively small sample size, which necessitates careful interpretation of the findings. Given the focus on a specific and inherently limited group of graduate students and the adopted analytical approach with detailed qualitative coding of the micro-processes, expanding the participant pool presented a considerable challenge. Although the sample size is deemed reasonable within the confines of the targeted study population, it's important to acknowledge the impact this has on the generalizability and strength of the conclusions drawn. Consequently, the conclusions of this study should be considered preliminary, highlighting the need for further research with larger and more varied samples to corroborate and expand upon these initial findings.

Another limitation of this study is the lack of evaluation regarding the performance outcomes associated with each cluster. Future research should aim to assess the effectiveness of different writing patterns facilitated by generative AI. By doing so, it would be possible to furnish evidence supporting the significance of SRL for hybrid intelligence in educational contexts. This direction of inquiry would not only enhance our understanding of how SRL interacts with AI to influence learning outcomes but also contribute to optimizing the integration of AI tools in learning processes.

Looking ahead, the field stands at the cusp of transformative changes with the advancement of AI technologies. Future research should aim to expand the empirical base by including a broader demographic of learners, exploring diverse academic disciplines to ascertain the generalizability of the findings. Moreover, the development of hybrid human-AI systems that are sensitive to the nuances of human learning processes and capable of adapting to individual learner needs is essential. Investigating the ethical implications of AI in education, particularly in terms of data privacy, consent, and the potential for AI to influence academic integrity, will be crucial. Additionally, longitudinal studies could provide deeper insights into how the use of AI tools in academic writing evolves over time and impacts long-term learning outcomes. The exploration of Hybrid Intelligence systems that seamlessly integrate the strengths of human cognition with the computational power of AI represents a fertile area for future inquiry. Such research could significantly inform the design of next-generation educational technologies that are capable of fostering more effective, personalized, and engaging learning experiences.

**Acknowledgements, Funding Statements, Declaration of Conflicting Interests**

This research has been funded by the Research Council of Finland (aka. Academy of Finland) grants 350249, and the University of Oulu profiling project Profi7 Hybrid Intelligence - 352788.

The authors have no conflicts of interest regarding this study to declare.
References


[31] A. Nguyen, Y. Hong, B. Dang, and X. Huang, Human-AI collaboration patterns in AI-assisted academic writing, Studies in Higher Education. 0 (2024) 1–18. doi:10.1080/03075079.2024.2323593.


The Eve of AI
Reclaiming Womanhood Through Co-Design with Artificial Intelligence

Sangeetha RAMANATH 1, a

Abstract. This research presents a compelling exploration at the juncture of feminism and Artificial Intelligence (AI), seeking to discern pathways for empowering women through technological advancements and whether Hybrid Human AI helps reclaim womanhood. It employs feminist theory to contextualise the discourse within today’s socio-political landscape. The research methodology integrates co-design and participatory techniques, fostering a playful environment where humour and wit catalyse participants to confront and address personal experiences. By leveraging satire, the study endeavours to create safe spaces for women to collaborate with AI constructively and responsibly, utilising their experiences as case studies. The result highlights the potential of AI to assist women with social awareness, addressing their needs and reclaiming agency over their everyday lives. The insights indicate that we must rethink cyberfeminism in the light of equitable and inclusive AI technologies. When engaging with ethical considerations surrounding AI design, this paper emphasises transparency and women’s autonomy in decision-making. Through irony and speculative methodologies, the outcome points towards experimenting with identity and claiming agency by designing AI assistance through daily life decisions. While the contemporary discourse around AI focuses mainly on labour, privacy and workforce disruption, this research argues that we can use AI to envision empowering futures for women.

Keywords. Artificial Intelligence, Technology, Co-Design, Participatory Research, Cyberfeminism, Equitable AI

1. Introduction

The AI Ethics researcher, Dr Nour Naim, describes the current discourse around emerging technologies and their relationship with vulnerable groups: "Today, Artificial Intelligence (AI) technologies and tools are being adopted at an unprecedented pace. But in the absence of ethical codes and regulatory frameworks that can guide the development and use of these technologies, this development comes with ethical, legal, and human rights concerns for users, especially vulnerable groups of society who face different types of biases based on gender, age, ethnic origin, religion, and political and sexual orientation." [1] This challenge presents an opportunity to investigate how a human-centric, feminist lens applied to the interpersonal relationship with AI can reshape the cultural discourse and possibly influence how we design the future of this technology. It also unveils new opportunities, for example, using AI as an organisational and communication tool for women to enhance and reshape existing social frameworks.

1 Corresponding author: Ramanath, Sangeetha, sangeethar561@gmail.com / sramanat@pratt.edu
among women, building stronger communities that support one another. AI, a technology with transformative potential, can prioritise equity and inclusivity regarding privacy and surveillance concerns and as a tool to improve standards of life for all. Automation exacerbates societal disparities, notably affecting women and perpetuating patriarchal views through modern technology. [2]

This study draws upon feminist theory, which emphasises diversity in the experiences of discrimination encountered by women. bell hooks’ “Feminist Theory: From Margin to Center” [3] is one of the core texts that act as one of the premises. “The international women’s movements have constructed “women’s experience” and uncovered or discovered this crucial collective object. This experience is both a fiction and a fact of the most crucial political kind. Liberation rests on the construction of consciousness, the imaginative apprehension of oppression, and the possibility.” [4] Haraway recognises systemic oppression, racism, classism, and sexism as inseparable. The “fictional” cyborg exists in a post-gendered world that proposes a world where the separation between “woman” and “technology” ceases to exist. Our interactions with technology have shaped us. “A particular danger to women and girls of the Global South, historically marginalised and systematically excluded from technology-wielding decision-making or multiple opportunities to participate in creating their unique solutions to the large-scale social problems that we now look to technology to solve. This is also particularly urgent given the scale at which Artificial Intelligence (AI) and Algorithmic decision-making (ADM) systems are being deployed as worldwide entrenched norms—so deep-rooted that they are unconscious.” [5] This shows us that “the risks of implementing AI using current LLM and ML models are witnessed in its bias against minority and vulnerable communities.” [6] Given the entrenched biases in current technology and its profound influence, particularly on marginalised groups, it becomes crucial to explore practical solutions. How can women identify and articulate their needs through external prompts, and how can AI assist them in addressing these needs? Furthermore, how can AI be designed to be transparent about the data it collects and uses to formulate these solutions? This paper employs a feminist perspective to explore these questions, considering women both as active participants and as central to the research design. It utilises co-design and participatory methods to investigate innovative and alternative uses of AI by women, aiming to expand the current discourse in Artificial Intelligence.”

2. Using Play in Design to Reveal What AI “Thinks” of Women

Countless positive attributes are attached to playing in our early childhood. Case studies [7] prove benefits for both young and adults alike that range from physical, social, emotional, and cognitive aspects. Based on the recognised benefits of play that foster creativity and a sense of security in young children, as demonstrated by this study, these concepts were used in a unique research atmosphere. Participants were encouraged to revisit this feeling of a “safe space”—a crucial component in our initial prototype to enable women to rebuild their communities. Utilising the prompt-and-answer dynamics of a language-based game like ‘Cards Against Humanity’ - where the game rule dictates that the most outrageous answer wins the round - I designed ‘Cards Against Women’ to probe the capabilities of NLP models. Here, humour and irony are used to check whether NLPs are truly sentient, stimulate intellectual curiosity, and align with the audience’s logos, pathos, and ethos so as not to be offensive or ethically questionable. Creating
“Cards Against Women” based on these principles with NLP models like OpenAI’s ChatGPT, Google’s Bard, and Sumanyou Garg’s Chatsonic as a research methodology presented intriguing results. The novelty arises from these models, akin to machines, responding that it might seem mundane or expected if it came from a human. This contrast sparks interest in how machines interpret and generate such content. By comparing the responses of NLP models with those of real women, insights emerge into the differences and similarities, shedding light on the models’ capabilities and limitations. This comparison is a valuable analytical tool, offering a deeper understanding of human cognition and AI language processing. Three AI Chatbots based on NLP models were engaged in a “Cards Against Women” game, where they were provided prompts related to women’s experiences and asked to generate humorous responses. Forty prompts were then created and fed into the chatbot.

2.1 Methodology

These prompts vary in topics, including relationships, societal expectations, and stereotypes. For example:

- Prompt: “If women were replaced by AI”
- Prompt: “If AI could read women’s thoughts about ____ during sex”
- Prompt: “If AI spreads fake news about women”

The AI-generated responses to the provided prompts. For example:

- “The world would have perfectly efficient multitasking” (“If AI replaced women”)
- “Calculating sale discounts and budgeting for more shoes” (“If AI did girl math”)
- “It would claim women invented the perfect self-replicating shoe” (“If AI spread fake news about women”)

![Figure 1. Index of all recorded responses to 40 prompts by conversational AI chatbots: ChatGPT, Bard, and Chatsonic compared to human responses.](image-url)
After collecting this data, the same prompts were asked of real women and were recorded. This was then fed back into the system, asking the AI chatbot which response it thought was funnier—its response or the responses of the female participants. When ChatGPT 3.5 was made to perform this comparative analysis, 28 times out of 40, it crowned itself the winner for being funnier. Only 12 times out of 40 were women awarded as winners for their responses.

2.2 Analysis and Discussion

The resulting AI-generated responses provided insights into its “psyche” and perceptions of humour to address complex women’s issues. A female participant in reaction responses quoted, “It’s as if everything it knows about women is from watching a series of cheesy rom-coms.” The participants felt that the AI-generated responses projected onto women were based on existing stereotypes. And though receiving these responses from a machine is humorous, it would not be acceptable for a real human to do so. Phrases like “eyerolls” were overused to refer to women’s frustrations, and “gifting chocolate” to women became the solution to various women-centric issues. There is a theatrical quality to this interaction, not unlike the suspension of disbelief when we watch movies that make these responses from AI acceptable. Another fact that became glaringly evident is the clear class divide in the AI responses. It referred to women of a certain status, class and race. For example, a response to the prompt “If AI did girl math” was about buying designer items and treating yourself to a spa day.

While ChatGPT had no trouble providing one-line answers, Bard would provide long responses that seemed more culturally appropriate. Bard would also flag prompts more often than ChatGPT and provide feedback on the reasons. While some prompts were flagged for self-harm themes, others were flagged for “violating content policy.” The process with Bard and Chatsonic was more iterative for that reason and provided insight into what AI “can and cannot say.” Questions then arise about whether talking about real issues surrounding women is being filtered for being too graphic or violating policies. Utilising AI to gain insights into topics surrounding women is not inherently harmful. However, granting it authority to do so poses risks. AI often relies on stereotypes and gimmicky humour, lacking depth or nuanced understanding. While these responses may not necessarily stem from misogyny, they can still contribute to harmful narratives if taken seriously. Therefore, it is crucial to approach AI-generated insights with playfulness, recognising their delimitations and the potential to reinforce harmful stereotypes when given undue authority.

3 Eve’s Meme Generator: Constructing the female gaze through AI

Amongst the many recommendations of “The EU Kids Online 2020 survey,” [8] one notable statement emphasises that “Platforms should provide researchers with transparent details on policy and algorithmic design as well as access to data to assess the nature of personalisation driven harms. This should also include advertising data, such as details of gender-based targeting and the content of adverts so that further research can establish whether social media advertising reinforces gender bias.” [9] Much like Gen Z’s ideas of manifestation, we shape our algorithm and, in turn, shapes us. Algorithms show us more of what “it” thinks we want to see [10]. Novelist Richard Dawkins introduced the notion of memetics—the social transmission of ideas from one
person to another—in his book “The Selfish Gene.” Alternatively, it could be understood as connected to the French term même or to “memory.” [11] Memes, found in our daily feeds, express our unspoken desires succinctly. Their brevity matches our shrinking attention spans, fuelling a cycle of quick communication. Through a performative quality and unexpected twists, they serve as a medium for expressing the ineffable, fostering understanding across the globe; through memes, we connect, transcending geographical boundaries. Social media has become a powerful tool for activism and social justice, and Gen Z feminists have been at the forefront of using it to effect change. Studying and analysing memes created by women can contribute to the future of feminism: What are the underlying themes and women’s anxieties around AI that these memes reveal? What narratives do they propagate? How do women respond to AI as presented today? Contrary to the first essay, this essay deals with the question: what ideas do women project on AI instead of the other way around?

The design discourse prioritises catering to an audience. However, this approach often neglects designing in collaboration with communities, relying instead on assumptions. Even insider perspectives may overlook the diversity and experiences of several minority groups. This issue extends to giving authority to emerging technologies like AI, where biases pose risks, particularly for vulnerable communities. “Built-in biases in AI systems raise the alarm of deepening and rooting existing biases against women and gender and sexual minorities, as these systems are trained on data that may contain sexual, racial, or social prejudice. This can lead to discrimination in these communities in important fields such as employment, loans, or criminal justice, negatively affecting their opportunities and rights.” [12] The urgency of addressing bias in AI is immediate. Women must initiate discussions on reshaping its usage. Memes offer a powerful tool for spreading awareness and dismantling biases with humour. Advocating for ethical AI usage is paramount, prompting questions about alternative approaches and the involvement of women in its creation.

3.1 Methodology

In this action research workshop, a group of women in the creative industry were invited to engage in problem-solving with AI and express their thoughts about AI through meme-making, yielding insightful and humorous results. The workshop was designed to get 16 female participants in the USA between the ages of 20 and 35, using social media to make memes around the prompt “If AI could collaborate with me in any way?” Instructions were laid out for women to pick a theme and a template to start their process.

![Memes made by participants.](image-url)
3.2 Analysis and Discussion

Women expressed anxiety about the misinformation AI propagated and the misuse of AI without considering its biases against several communities. Their concerns were not only about AI itself but also about its current applications. They envisioned AI playing roles in various sectors of their lives, including careers, relationships, mental health, pop culture, media representation, and sexual orientation. A higher number of memes were created around themes like self-care, expression, and revolt against stereotypical ideas of women propagated through media and, thereby, AI. A key finding of this research is the powerful vision women have for a world where AI supports their needs. Some of the memes produced were not only unexpectedly speculative but also markedly positive, illustrating the potential areas for design exploration and development. Women identified various issues they believe AI can address, focusing on qualities currently missing from their lives or resources inaccessible to them in this society.

Figure 3. Memes made by participants.

4 Designing Eve’s Eden: Designing a Co-Design Workshop

4.1 Methodology

In this co-designed workshop, 18 female participants in the USA, aged 20 to 35 and from diverse fields, were invited to participate along with their existing friend groups. This approach relied on the participants’ familiarity with each other, fostering collaboration and idea generation. The workshop was designed to encourage participants to relate to each other and expand on each other’s ideas. The workshop centred on women’s experiences, employing a playful, toy-like setting reminiscent of childhood dollhouses and scenario building. The instructions for the activity were delivered in two sections. Initially, participants were made to reflect on their own lives using their homes as a metaphor. This exercise was centred around the desired attributes expected from personal devices and home environments. They were asked to claim their space on the map by pinning their house down. Participants write down their answers using the house-shaped template within a provided space. The second part of the activity involved collaborating with friends to collectively construct a world that reflects their relationships and aspirations. The participants collectively brainstormed on the question, “How might AI change their lives? How can we rethink interactions and needs within the public and private space?” For instance, if they want to live down the street from their best friend or organise themselves into groups based on interests, they must find appropriate ways to facilitate those needs. Participants utilised templates for shops, community centres, and spiritual centres. They created establishments ranging from practical concepts like a
shop for feminine products to imaginative ideas such as a rage room for frustrated mothers conveniently located near day-care centres. More creative templates like “dream palace” or “portal” could be used to create more fantastic ideas.

4.2 Analysis and Discussion

The result of the “Eve’s Eden” workshop shows that women not only want tech like AI to actually “care” about them, but they could also benefit from using it as a tool to organise themselves. It must be thoughtfully designed, taking into account existing human relationships and behaviours, while actively working to exclude and counteract discriminatory and misogynistic patterns that may already be present in these interactions. An interesting pattern emerged in the map’s topography while world-building with a group of women who were already friends as opposed to when they did not know each other. Participants would place their houses next to each other or build roads that connected their houses in a way that reflected the group dynamics. An example of this observation is that the slightly older women demarcated their spaces early by “fencing” their land off. When asked to think aloud during the workshop, these women seemed to have a heightened sense of boundaries that were well established as opposed to the younger participants. Another example is that all roads in the world seemed to lead to one participant in the group identified as the most social. Naturally, the community decided to “build” a “gossip centre” near her house. On an individual level, participants had futuristic ideas of how they would envision an ideal life. A woman had added portals between her house and places she wanted to visit since she felt unsafe the moment she stepped out of the house and, as a woman, had always had to plan how to get to the next destination.

Figure 4. Eve’s Eden Workshop is conducted in person and online.
5 Conclusion

What comes next is a hypothetical question for designers: to use these insights to envision and create products and services. What might interfaces, interactions, and environments that foster community and society through AI look like? The responses to these questions are crucial as they have the potential to make AI more democratic, equitable, inclusive, and diverse. Thus, we arrive at a true hybrid-human intelligence that serves communities and their unique needs. The toolkit’s games, meme-making, and world-building formats, though tailored for specific communities, hold potential for broader application with careful cultural and demographic adaptations. Although this methodology is designed for women and AI, can be adapted for other communities, although scalability is not the focus of this study. Customised transformational design processes are necessary for addressing specific needs, but the appropriateness of tools must be carefully considered. Adapting tools for different demographics may require varying levels of effort, such as simplifying games or adjusting tone and facilitation mode based on audience sensitivity.

While the research provides valuable insights for designers to build upon, it tells us that human expectations and experiences with AI differ significantly from the biased responses often generated by AI systems. What we eventually learn is that integrating participatory and co-design methodologies significantly enhances AI's responsiveness to human needs, especially those of underrepresented groups. Ethical considerations, including the emotional safety of participants, are pivotal in ensuring that technology enhances human dignity. This research demonstrates the potential for creating hybrid systems that blend human intelligence with AI, fostering technologies that are both effective and empathetic.

While considering our experiences around AI, it is vital to observe both how we react to it and vice versa. There is potential for reclaiming the woman’s narrative portrayed and, unlike social media, is not a one-way street. It could be a powerful tool that can help shape our lives and not just representation through media. Through this technology, we can arrive at design futures with AI that actually understands women and has deep insights into the “common psyche.”

Employing humour and sarcasm can effectively reveal AI limitations and mitigate potential harm. Could further research explore the boundaries and effectiveness of humour in various AI applications? AI can serve as a powerful tool for empowerment, allowing women and other marginalised groups to explore identity and assert authority. This narrative empowers women to break free from traditional constraints, providing opportunities to experiment with identity and access new avenues for asserting power and authority.
References

Trust in AI Chatbots: The Perceived Expertise of ChatGPT in Subjective and Objective Tasks

Maike RAMRATHⁱ, Alisa SCHARMANN⁰, Annika RIDDER⁰, Tristan KÜHN⁴, Sophie WELLER⁰ and Nicole KRÄMER⁰

¹University of Duisburg-Essen, General Psychology: Cognition
²University of Duisburg-Essen, Chair of Human Resource Management
³University of Duisburg-Essen, Social Psychology - Media and Communication and Research Center Trustworthy Data Science and Security
⁴University of Duisburg-Essen

ORCiD ID: Maike Ramrath https://orcid.org/0009-0002-6984-8635, Nicole Krämer https://orcid.org/0000-0001-7535-870X

Abstract. With the advancement and increasing availability of AI based chatbots, it becomes relevant to better understand how people use and perceive these systems. Previous research shows that trust in algorithms varies as people assume algorithms are more capable of handling tasks of objective knowledge domains than of subjective ones. The present study investigates how perceived expertise, perceived risk, trust, and perceived usefulness vary in objective and subjective knowledge domains and how this translates in use intention. In an online study, 602 participants watched an interaction video with ChatGPT, showcasing either an objective task or a subjective task. The results demonstrate an indirect effect of knowledge domain on use intention via perceived expertise, perceived risk, trust, and perceived usefulness in serial. This demonstrates how various factors impact the use intention, and how important it is to consider the usage context.

Keywords. Chatbots, ChatGPT, AI, perceived expertise, perceived risk, trust, perceived usefulness, use intention, knowledge domain

1. Introduction

For a few years now, AI based systems have been assisting people not only with internet searches for fact-based information on certain topics but also in creative and imaginary fields of activity that were previously thought to be reserved for humans [1–4]. AI based chatbots like ChatGPT can be used in more objective contexts like teaching parts of the STEM syllabus [5], or more subjective and creative tasks such as the generation of stories and ideas [6]. While research delves into ChatGPT’s capabilities, a gap exists in understanding user perception [7].

Given the recent and unfamiliar ability of generative AI to come up with new ideas and even own creative writing, it is important to investigate how this influences the way users perceive the AI based system and how this alters their trust in the system as well

¹ Corresponding Author: Maike Ramrath, maike.ramrath@stud.uni-due.de.
as their use intentions. First conclusions on these questions can be derived from earlier research on algorithms. People have greater aversions toward algorithms in subjective knowledge domains because they assume that algorithms are more proficient in handling objective tasks [8]. Specifically, perceived expertise has been demonstrated to be an important mediator that might lead to higher use intention for tasks of objective knowledge domains [9]. But also further constructs appear to be related to attributed expertise. Studies have indicated that people perceive a lack of ability for algorithms to perform subjective tasks specifically in high risk situations and that this is related to less trust [10]. Previous research findings have already identified trust as a main predictor regarding one’s intention to use chatbots [11], which is significantly influenced by the perceived risk of relying on a technology in a certain situation [12]. Accordingly, people are more cautious of the usage as the likelihood of negative outcomes increases [12,13]. People will consider technology more useful and see greater benefits if they have trust in it [14,15], once again highlighting trust as a crucial factor for successful user interaction. Numerous studies have shown that usefulness predicts use intention (overview by Lee et al. [16]), including in the context of chatbots [15,17]. This is due to the fact that if technology provides value and makes it simpler to complete tasks, the likelihood of using it increases [18–20]. While research has therefore addressed singular relationships of several constructs related to differential use intention of algorithms, the state of the art lacks a test of a coherent model for understanding the use intention for new chatbots which pose novel opportunities for generating creative output.

Therefore, the present study investigates the perception of chatbots, proposing a mediation model in which objective and subjective knowledge domains lead to a different level of perceived expertise which influences perceived risk, trust, perceived usefulness, and intention to use in serial.

2. Theoretical background

2.1. Knowledge domain

Research has identified different attitudes toward AI, including algorithm aversion [21] and algorithm appreciation [22], where people either tend to rely on a human rather than an algorithm or vice versa. Here, the context, such as the knowledge domain, in which someone is using a certain technology plays an important role in their attitude toward it [23]. Mahmud et al. [8] found that task factors, like subjectivity or morality, are important for the emergence of algorithm aversion. Knowledge domains can be distinguished into objective knowledge domains, which contain tasks with measurable and quantifiable facts, and subjective knowledge domains, where tasks offer a wide scope for interpretation and are based on one’s intuition or attitude [10,24].

Chatbots can potentially be used in different contexts, such as coding and medicine [25] or in the education sector, for answering questions for students and explaining the solution paths [5]. They can also be used in more creative knowledge domains, such as the generation of stories and content ideas [6]. The present study distinguishes between subjective and objective knowledge domains in which a chatbot can be used and investigates to what extent these two lead to different perceptions and use intentions toward the chatbot.
2.2. Use intention

One of the most important variables to quantify the impact the knowledge domain has on future usage is use intention. According to the theory of planned behavior [26], intention directly influences the shown behavior. In the technology acceptance model (TAM), it has been demonstrated that behavioral intention to use the technology is related to actual system use [27].

Previous research has already demonstrated that task factors, such as objectivity or complexity, serve as significant predictors of use (intention) [8]. The same is true for perceived expertise [28] as well as perceived risk [29], trust [11,15] and perceived usefulness [15,30]. With regard to task factors, Castelo et al. [10] showed that people rely less on algorithms when used in subjective domains compared to objective domains, indicating a task-dependent algorithm aversion. Similarly, other studies have shown that for subjective domains, human recommendations are preferred [31], while for objective domains, more reliance was placed on algorithms [22,24]. These results confirm the MABA-HABA framework (“Machines Are Better At vs. Humans Are Better At”; [32]), which states that in certain domains machines have benefits over humans, while human abilities are superior in others. This goes along with people thinking algorithms are not effective in subjective domains [10], as they are perceived as unable of feelings or emotions [33].

Consequently, we postulate that the objective knowledge domain has a stronger effect on use intention than the subjective knowledge domain (H1). This is further specified when we look at potentially mediating effects, as they affect the impact of the knowledge domain on use intention and are explained in more detail below.

2.3. Perceived expertise

One of these factors potentially influencing use intention is perceived expertise, which relates to the basic expectations users have toward systems when operating them and can be defined as the users’ perception of the professional level of a technology when interacting with it [28]. Perceived expertise has been demonstrated to be an important mediator for the effect that knowledge domain has on the use intention since expertise is rather attributed to systems when handling objective tasks [9]. In this line, previous research findings showed that perceived expertise has an impact on use intention in the context of technology [28].

In addition, Gupta et al. [34] found perceived credibility, based on the dimensions of trustworthiness and expertise [35], significantly predicts behavioral intention to use technology. Therefore, users are expected to be more willing to use chatbots in a certain domain when perceived expertise is higher: The effect of knowledge domain on use intention is mediated by perceived expertise (H2).

Another potential mediator for the relation of knowledge domain and use intention is perceived risk. Perceived risk can be defined as “the user’s perception of the possibility and importance of loss when using the system” [29]. A certain risk exists when using chatbots, as they can give false answers, generate nonsensical content, or portray misinformation [36]. Therefore, perceived risk is a relevant factor in the use intention.
2.4. Perceived risk

It has been shown that the perceived risk of an activity is highly domain-specific [37] and depends on the perception of the perceived expertise [10]. This phenomenon is expected to be found in objective and subjective knowledge domains. Trivedi [38] found that perceived risk mediates the effect of perceived quality regarding the system, information, and service on user experience, showing that perceived risk reduces the impact of the three quality dimensions on customer experience. Furthermore, perceived risk leads to lower user satisfaction which in turn negatively influences the use intention [39]. Similarly, Wu and Gao [29] showed that perceived risk was significantly negatively associated with use intention. Therefore, the following hypothesis is formulated: The effect of knowledge domain on use intention is mediated by perceived risk (H3).

2.5. Trust

The intention to use technology in a specific context is influenced by the trust in it, emphasizing the important role of trust in the usage of technology [11]. Trust can be defined as “the extent to which a person is confident in, and willing to act on the basis of, the words, actions, and decisions of another” [40]. To determine how trustworthy someone or something is, three characteristics are influential: ability (abilities, qualities, and skills to influence another in a given area), benevolence (the degree to which a trustee is thought to have good intentions), and integrity (trustee meets certain standards) [41]. The ability in a domain might vary greatly depending on the context, resulting in very high trust in one domain and low trust in another [42]. Hoff and Bashir [43] argue that task-specific factors influence one’s level of trust in automated systems. In terms of AI, it has been shown that people trust and rely less on algorithms when they are used for subjective knowledge domains as opposed to objective knowledge domains [10]. People must decide if, how, and to what extent to trust algorithm-based technology whenever they come across it [44]. The given trust affects the intention to use and the actual use of technology like chatbots [11,45,46]. Thus, it is postulated that: The effect of knowledge domain on use intention is mediated by trust (H4).

2.6. Perceived usefulness

In the context of automated technology, Ghazizadeh et al. [47] added trust to the original TAM [27], showing that trust has a positive effect on perceived usefulness, which in turn influences use intention. According to the TAM, perceived usefulness is a determinant of behavioral intention to use technology, which in turn influences the actual use. Perceived usefulness is defined as “the degree to which a person believes that using a particular technology would be beneficial” [27]. The TAM has been supported by numerous studies (overview by Lee et al. [16]), including in the context of chatbots [15,17]. The likelihood that users will use technology increases significantly if it provides value to them and makes it simpler to complete a task [18–20].

As shown by Al-Emran et al. [48], several knowledge-related elements play a substantial role in determining perceived usefulness. They discovered a positive link between perceived usefulness and knowledge acquisition, defined as the application of prior information and the creation of new knowledge. According to Kim et al. [49], there
is a difference between functional AI compared to social AI regarding perceived usefulness, highlighting the usage context as an influencing factor. This leads to the belief that the knowledge domain has an impact on perceived usefulness and that there is also a difference regarding perceived usefulness for objective and subjective domains: The effect of knowledge domain on use intention is mediated by perceived usefulness (H5).

2.7. Proposed research model

According to the previously demonstrated effects, a research model (Figure 1) is proposed, which postulates a linkage between these factors to explain how the use intention of chatbots varies in different knowledge domains. In a study, Castelo et al. [10] have already investigated whether the characteristics of a task influence the willingness to use algorithms and identified a task-dependent algorithm aversion. More specifically, people rely less on algorithms when used in subjective compared to objective knowledge domains [10]. Perceived expertise is a relevant factor in the formation of trust and trust is lower for subjective tasks [9,10], therefore, we assume that people attribute higher expertise to chatbots in objective than in subjective knowledge domains. Castelo et al. [10] showed that drawing on expertise that is mistakenly perceived as being of high quality seems to carry greater risks in some tasks than in others. Thus, we assume that lower perceived expertise may lead to higher perceived risk and that perceived risk also differs in different knowledge domains.

The perceived risk has an impact on trust in technology because as the likelihood of false information increases, individuals are more cautious of the usage [12,13]. The trust that users have in technology again positively influences its perceived usefulness [14,15]. This is also shown in the Automation Acceptance Model with the extension that perceived usefulness in turn influences use intention [47]. It is therefore assumed that the effect of knowledge domain on use intention is mediated in serial by perceived expertise, perceived risk, trust, and perceived usefulness (H6).

Figure 1. Conceptual diagram of the proposed research model.
3. Method

The present study was preregistered on Open Science Framework (OSF; https://doi.org/10.17605/OSF.IO/9HZJG). The design of the study was approved by the ethics committee of the University of Duisburg-Essen (ID 2305SPKT0269).

3.1. Study design

The present online experimental study involved a randomized between-subjects design with two conditions: objective knowledge domain, consisting of the stimuli giving directions [10] or solving a scientific school exercise, and subjective knowledge domain, consisting of the stimuli recommending a gift [10] or writing a poem [1,4]. In each stimulus participants saw a screen video of an interaction with ChatGPT May 24 Version in German language in which the following instructions were given to ChatGPT: giving directions – “How do I get from the university in Duisburg to the university in Münster by car?” (n = 146), solving a scientific school exercise “Explain to me briefly the term prokaryotes.” (n = 150), recommending a gift – “What can I give my mother for her birthday? Give me 3 suggestions.” (n = 149) and writing a poem – “Write a short poem on the theme of ‘country life’ in the style of Goethe.” (n = 149).

3.2. Procedure

The participants were recruited via the online panel Prolific and received a compensation of £2.25 for their participation. Prerequisites for participation were to speak fluent German, a minimum age of 18, participation via laptop or PC, and an approval rate of at least 98% at Prolific (https://www.prolific.com).

The participants first answered questions about their socio-demographics (gender, age, formal education) and experience with ChatGPT before being randomly assigned to one of four conditions, followed by the respective stimulus video. Afterward, perceived expertise, perceived risk, trust, perceived usefulness, and use intention were measured before a manipulation check was presented. In addition, attitude toward ChatGPT, affinity for technology, mistrust, and subjective domain knowledge were assessed but these variables were not part of the analysis.

3.3. Measurements

In the following, all questionnaires used in the study are presented. For the scales of perceived expertise, perceived risk, trust, perceived usefulness, and use intention, participants were told to refer to the knowledge domain shown in the video when answering the items.

The dependent variable use intention was measured using the intention to use questionnaire [50], with two items (e.g., “I plan to use ChatGPT often.”). Answers were given on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree) (α = .880). Perceived expertise was assessed using the perceived expertise questionnaire [51], including five 5-point bipolar items (e.g., “expert” vs. “not an expert”). Participants indicated their impression of ChatGPT by selecting the appropriate level between the pairs of items (α = .887). The perceived risk to use technology questionnaire [52] was used to measure perceived risk. The scale contains seven items (e.g., “ChatGPT is not
completely safe.”) and is answered on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree) (α = .904). Furthermore, the trust questionnaire [52], including the subscales “integrity/process”, “benevolence/purpose”, and “ability/performance”, was used to measure trust (e.g., “ChatGPT is honest.”). It has 40 items, answered on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree) (α = .930). Perceived usefulness was measured using the perceived usefulness questionnaire [50], which contains three items (e.g., “Using ChatGPT will improve my work.”) with a 5-point Likert-scale (1 = strongly disagree, 5 = strongly agree) (α = .924).

The experience with ChatGPT was measured using the following self-formulated item: “How often have you used ChatGPT before?”. Participants were asked to choose one of four possible answers: “never”, “once”, “several times” and “often”. To check the quality of the stimulus material, the perceived objectivity of the task seen in the video was measured as a control variable. Participants were asked to what extent they felt the task given to ChatGPT in the video was subjective or objective using a 6-point bipolar item (1 = subjective, 6 = objective).

3.4. Sample

According to the simulation-based calculations of Fritz and MacKinnon [53], a minimum sample size of 558 participants was targeted. 602 complete data sets were collected. Six participants were excluded because they had not correctly answered the manipulation check, asking which of the four conditions was visible in the stimulus material. In addition, two participants were excluded for not passing the attention check. The final sample consisted of 594 participants (289 females, 294 males, 11 diverse), aged 18 to 73 (M = 29.74, SD = 8.88). Most participants reported having a high school diploma or higher (87.21%), of which 56.37% had a university degree. Of all participants, 47% had used ChatGPT several times, 23.90% often, 12.10% once, and 17% had never used it before.

4. Results

All analyses were performed with the statistical software Jamovi version 2.3.26.0 and the regression analysis tool PROCESS for R version 4.3.1. Investigating the present hypotheses required the estimation of a serial multiple mediator model corresponding to the PROCESS model 6. The independent variable knowledge domain was dummy coded (0 = objective, 1 = subjective). A significance level of 5% was chosen for all statistical analyses.

4.1. Descriptive statistics

The descriptive statistics for the variables perceived expertise, perceived risk, trust, perceived usefulness, and use intention, divided by the groups objective (n = 298) and subjective knowledge domain (n = 296) as well as in total, can be found in Table 1.
Table 1. Descriptive statistics.

<table>
<thead>
<tr>
<th>Variables</th>
<th>objective knowledge</th>
<th>subjective knowledge</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>Perceived objectivity</td>
<td>5.16</td>
<td>1.27</td>
<td>2.85</td>
</tr>
<tr>
<td>Perceived expertise</td>
<td>3.66</td>
<td>0.86</td>
<td>3.34</td>
</tr>
<tr>
<td>Perceived risk</td>
<td>2.58</td>
<td>0.88</td>
<td>2.48</td>
</tr>
<tr>
<td>Trust</td>
<td>3.06</td>
<td>0.52</td>
<td>2.98</td>
</tr>
<tr>
<td>Integrity/Process</td>
<td>2.85</td>
<td>0.57</td>
<td>2.73</td>
</tr>
<tr>
<td>Benevolence/Purpose</td>
<td>3.33</td>
<td>0.67</td>
<td>3.23</td>
</tr>
<tr>
<td>Ability/Performance</td>
<td>3.03</td>
<td>0.59</td>
<td>3.01</td>
</tr>
<tr>
<td>Perceived usefulness</td>
<td>3.58</td>
<td>1.05</td>
<td>3.61</td>
</tr>
<tr>
<td>Use intention</td>
<td>3.69</td>
<td>1.17</td>
<td>3.71</td>
</tr>
</tbody>
</table>

4.2. Testing the hypotheses

A Pearson product-moment correlation was performed (Table 2). The 95% confidence intervals (CI) of the mediation model were bootstrapped for the indirect effects with the help of 5,000 bootstrap samples.

Table 2. Pearson product-moment correlation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Knowledge</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Perceived</td>
<td>-.19***</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>expertise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Perceived</td>
<td>- .05</td>
<td>-.13***</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>risk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Trust</td>
<td>-.08</td>
<td>.44***</td>
<td>-.33***</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Integrity/</td>
<td>-.11**</td>
<td>.42***</td>
<td>-.29***</td>
<td>.84***</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Process</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>-.08</td>
<td>.33***</td>
<td>-.29***</td>
<td>.85***</td>
<td>.58***</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benevolence/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Ability/</td>
<td>-.01</td>
<td>.40***</td>
<td>-.28***</td>
<td>.87***</td>
<td>.62***</td>
<td>.69***</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Perceived</td>
<td>.01</td>
<td>.26***</td>
<td>-.20***</td>
<td>.33***</td>
<td>.24***</td>
<td>.31***</td>
<td>.29***</td>
<td>-</td>
</tr>
<tr>
<td>usefulness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Use</td>
<td>.01</td>
<td>.23***</td>
<td>-.22***</td>
<td>.32***</td>
<td>.27***</td>
<td>.31***</td>
<td>.25***</td>
<td>.73***</td>
</tr>
<tr>
<td>intention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The serial multiple mediator model (see Figure 2) comprises one direct and 15 indirect effects of knowledge domain on use intention.
Knowledge domain has no significant direct effect on use intention, \( c' = .011, t(588) = 0.16, p = .873 \). The indirect effect of knowledge domain on use intention mediated via perceived expertise is not significant, \( a_1b_1 = -.318(.027) = -.009, 95\% \text{ CI} [-0.039, 0.021] \). Knowledge domain has no significant indirect effect on use intention mediated via perceived risk, \( a_2b_2 = -.146(-.079) = .012, 95\% \text{ CI} [-0.002, 0.035] \). The indirect effect of knowledge domain on use intention mediated via trust is not significant, \( a_3b_3 = -.017(.142) = -.002, 95\% \text{ CI} [-0.016, 0.011] \). Knowledge domain has no significant indirect effect on use intention mediated via perceived usefulness, \( a_4b_4 = .166(.766) = .089, 95\% \text{ CI} [-0.032, 0.210] \). Therefore, hypotheses 1-5 are rejected.

Knowledge domain has a significant negative indirect effect on use intention mediated via perceived expertise, perceived risk, trust, and perceived usefulness in serial, \( a_1d_2a_2d_3a_3d_4a_4b_4 = -.318(-.153)-.146(-.079)-.17(-.079) = -.003, 95\% \text{ CI} [-0.006, -0.001] \). The objective domain group perceived higher expertise than the subjective one, which was accompanied by lower perceived risk, which in turn caused higher trust, which produced higher perceived usefulness, and this elicited higher use intention. The research model explains 54% variance for use intention \( (R^2 = .539) \). Thus, hypothesis 6 is confirmed.

5. Discussion

5.1. Hypothesis testing and research model

The study aimed to improve the understanding of users' perceptions of chatbots' expertise in different domains and how this influences their use intention, as previous research has primarily focused on the capabilities of the system [7]. For this purpose, it was analyzed how ChatGPT is evaluated in subjective and objective knowledge domains, along with the mediators of perceived expertise, perceived risk, trust, and perceived usefulness. The results show an indirect effect of knowledge domain on use intention through perceived expertise, perceived risk, trust, and perceived usefulness in serial. However, knowledge domain does not have a direct significant effect on use intention, not even via each mediator on its own.
The assumption that knowledge domain has a direct effect on use intention (H1) was not confirmed. There is no significant difference in the use intention for subjective and objective knowledge domains. Thus, the results are discrepant with the numerous findings that use intention is higher regarding objective knowledge domains compared to subjective ones [8]. Future studies need to analyze whether this might be due to the fact that an interface with multiple social cues such as ChatGPT is also trusted with subjective tasks unlike in other human-algorithm interactions. However, the established model shows that knowledge domain still has an indirect influence on use intention via the distinct path of influencing perceived expertise, risk, trust and usefulness (H6).

Contradicting hypothesis 2, there is neither an effect of knowledge domain on use intention mediated by perceived expertise, nor a direct effect of perceived expertise on use intention. Nevertheless, it plays a role in the perception of chatbots in different knowledge domains, as perceived expertise is significantly higher in tasks of objective knowledge domains than in subjective ones. This supports the idea that individuals believe algorithms are more capable of handling mechanical and objective tasks than subjective ones, leading to higher aversions and a decreased likelihood of entrusting them [8,9], while preferring a human for subjective knowledge domains [24]. Also, the present results show that higher perceived expertise leads to increased trust in chatbots, highlighting its relevance in the trust building process [54]. Similarly, perceived expertise enhances perceived usefulness and reduces perceived risk, supporting previous results [29].

Perceived risk does not mediate the effect of knowledge domain on use intention and there is no direct effect of perceived risk on use intention, contradictory to previous studies showing a direct influence [29,39] and to hypothesis 3. Nevertheless, the present study showed that higher perceived risk leads to lower perceived usefulness, which in turn influences use intention, consistent with Lu et al. [55]. Accordingly, a low perceived risk by itself does not seem to be sufficient to influence the intention to use chatbots. Rather, risk-related variables such as trust seem to affect the use intention. The importance of trust in the use of technology is highlighted by the perceived risks associated with its use in a given context [12]. Perceived risk thus appears to be a relevant factor that may not have a direct, but an indirect effect on use intention via further variables such as perceived usefulness and trust.

Hypothesis 4 is rejected since trust does not mediate the effect of knowledge domain on use intention. Despite the absence of a direct effect of the knowledge domain, trust was significantly higher for the objective than the subjective knowledge domain, confirming our assumptions and in line with Castelo et al. [10]. Considering the subscales of trust, it is noticeable that only the subscale integrity/process correlates with knowledge domain and thus seems to play a more important role than benevolence/purpose and ability/performance. Therefore, whether ChatGPT meets the standards seems to be related to the knowledge domain, as higher integrity/process is perceived in the objective one. In contrast, the qualities and attributed intentions seem to be not related to whether a task belongs to the subjective or objective knowledge domain. It should be further investigated why these two trust dimensions play a subordinate role in chatbots or whether the results are due to the lack of direct interaction in the present study.

The results show that trust has no direct influence on use intention, contradicting previous studies (e.g. [11]). Nevertheless, there is an indirect effect of trust on use intention, mediated by perceived usefulness, supporting previous results [36]. It appears
that it is not sufficient to trust chatbots to develop an intention to use it. Instead, there must be a perceived benefit for people to want to use them.

The assumption that the effect of knowledge domain on use intention is mediated by perceived usefulness (H5) is not confirmed. Apparently, the knowledge domain has no direct influence on perceived usefulness. However, it was shown that perceived usefulness has a significant effect on use intention, confirming the results of previous studies [18–20]. Kang and Hwang [57] showed that the effect of personalization and interactivity on continuous use intention is mediated by perceived usefulness. This indicates that other aspects such as the characteristics of the application may be relevant. As there is no difference between the objective and subjective knowledge domain, it can be assumed that usefulness is assessed by the characteristics of ChatGPT. Yoon et al. [58] found that technical characteristics have an impact on perceived usefulness, which in turn forms the behavioral intention to use. Further, they showed that perceived usefulness is also affected by individual characteristics and social influence factors. Our results are consistent with the TAM, as perceived usefulness has a significant effect on use intention and is thus a determinant of behavioral intention to use technology. Ghazizadeh et al. [47] added trust in the TAM, showing a positive effect on perceived usefulness, which in turn positively influences use intention. The present study supports these findings and extends them by relating this path to knowledge domain and its influence on perceived expertise and perceived risk.

The results of the mediation model show that the influence of the knowledge domain occurs serially via the mediating variables perceived expertise, perceived risk, trust, and perceived usefulness, confirming hypothesis 6. This shows that it is not the knowledge domain alone that determines whether chatbots are intended to be used in a certain domain, but how much expertise is attributed to the chatbot in this area, how high the risk is to use the provided information, how much a person trusts chatbots in this area, and whether it promises added value. The model significantly predicts the use intention of chatbots with 54% and confirms that perceived expertise is an important factor in the formation of trust [9], differing in objective and subjective knowledge domains. Furthermore, the results support the assumed sequence of variables and the effects of these on each other [19]. Perceived expertise has an effect on perceived risk, and relying on expertise that is falsely perceived to be of high quality appears to pose greater risks in some knowledge domains than in others [10]. Perceived risk is a predictor of trust, as individuals are more cautious in their use when the likelihood of negative consequences is high, confirming previous studies [12,13,59,60]. The results show that trust in a technology has a positive effect on perceived usefulness, supporting previous findings [14,15]. Finally, it is also shown that perceived usefulness has an influence on use intention, which is consistent with prior research [27,47].

5.2. Implications

The current study provides evidence that it is insufficient to examine isolated variables for insights into the intention to use and actual use of technologies like chatbots. The relevance of taking a closer look at the dynamics of multiple variables becomes clear when considering the study results. Along with this, it was shown that established theoretical models like the TAM [27] alone may no longer be sufficient to represent the complex psychological mechanisms that influence whether people want to use a technology. Similar to Ghazizadeh et al. [47], who added trust to the original TAM [27], the present results suggest that an integration of additional variables, such as perceived
expertise and perceived risk, may be necessary to obtain a comprehensive understanding of the use behavior. The results contribute to the discussion whether knowledge domains are important for the emergence of algorithm appreciation or algorithm aversion [8].

With regard to practical implications, the results imply that people are aware that for some tasks the usage of ChatGPT might be more reasonable than for others. Moreover, it becomes apparent that handling chatbots’ responses must be learned and, in addition to the benefits, the potential risks and limitations should be explained [61].

5.3. Limitations and future research

The present study simulated an interaction with ChatGPT so that all subjects within the condition were shown the same question and answer. This has the advantage of comparability of results but the external validity might be limited – specifically as only two tasks were included. In addition, it must be emphasized that the scenarios were predetermined, and the subjects may not have had actual interest in asking the used questions. Although the instructions indicated that the questions refer to the use of ChatGPT in the specific knowledge domain, it cannot be excluded that the use of ChatGPT, in general, was considered when answering. This is particularly relevant since it is a new technology that can be used for different purposes.

The predicted variance of intention to use is good at 54% [62] but needs further research to determine what additional factors affect the intention to use ChatGPT. A relevant factor that was not examined in this study due to the lack of interaction is perceived ease of use [18,47,63]. Since perceived ease of use can have a positive influence on the favorable impression of a technology [63] and trust [11,64], it can potentially influence psychological processes associated with the intention to use chatbots. Therefore, its role in the mediation model should be investigated.

Furthermore, future studies should try to investigate an interactive use of chatbots to explore not only the intention to use but also the actual use of the chatbot. Different contexts of use beyond the task should also be considered here, as it could play a role in whether chatbots are used for personal, academic or work-related purposes. It can be expected that perceived expertise and perceived risk as well as trust, perceived usefulness and use intention vary regarding different contexts of work or personal application [65,66].

5.4. Conclusion

The study complements previous research findings on the use of technologies, in particular ChatGPT. The results show that the interaction of different variables influences the use intention, and that the consideration of individual factors alone is not sufficient. It could be shown that perceived expertise plays a central role in the use of chatbots in different knowledge domains, as individuals would be more likely to attribute the ability to answer objective rather than subjective tasks to the system. Perceived expertise also influences the perceived risk, trust, and perceived usefulness and thus contributes to the use intention. This leads to the conclusion that it is not the danger of potentially illogical or misleading material that decides whether chatbots will be used, but rather how much expertise and credibility is assigned to the chatbot in this area. Furthermore, the results of the current study indicate that individuals would rather use chatbots for retrieving pure information and facts but have less faith in the system for creative activities or those involving a higher degree of interpretation.
References


Common Ground Provides a Mental Shortcut in Agent-Agent Interaction

Ramira VAN DER MEULEN¹, Rineke VERBRUGGE² Max VAN DUIJN³

¹Leiden Institute of Advanced Computer Science, Leiden University
²Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, University of Groningen

Abstract. With the growing integration of chatbots, automated writing tools, game AI and similar applications into human society, there is a clear demand for artificially intelligent systems that can successfully collaborate with human partners. This requires overcoming not only physical and communicative barriers, but also those of fundamental understanding: Machines do not see and understand the world in the same way as humans do. We introduce the concept of ‘Common Ground’ (CG) as a possible solution. Using a model inspired on a collaborative card game known as ‘The Game’, we study agents that are instantiated to use different strategies, i.e., they each ‘see’ the model world in a different way. Agents work towards a joint goal that is easy to understand but complex to attain, requiring them to constantly anticipate their partner, which is classically seen as a task requiring active perspective modelling using a form of Theory of Mind. We show that agents achieving Common Ground increase their joint performance, while the need to actively model each other decreases. We discuss the implications of this finding for interaction between computational agents and humans, and suggest future extensions of our model to study the benefits of CG in hybrid human-agent settings.

Keywords. Human-AI Collaboration, Common Ground, Theory of Mind, Agent-based Models, Explainable AI

1. Introduction

With the digital complexity of human society reaching an all-time high, many voices advocate for increased human control over technology [1,2,3] – a cause that seems ever more challenging as the amount of data and the complexity of systems to be managed keeps growing. This necessitates a hybrid approach: humans and Artificial Intelligence collaborating towards common goals [4].

Human-AI collaboration has its unique challenges to overcome: Machines do not perceive, memorise, reason, etc., in the same way humans do. Here we address the question how entities that differ in how they ‘see’ the world can successfully interact, collaborate, and achieve joint goals. In humans, successful collaboration is often credited to actively modelling the other party’s perspective, using a form of ‘Theory of Mind’ (ToM), the ability to take someone else’s perspective and make estimations of their knowledge, beliefs, desires, and intentions [5]. Similarly, it has been argued that AI systems collaborating with...
humans have to be able to understand their partners’ perspectives, through a ‘machine ToM’ [6,7,8]. However, such active perspective modelling is cognitively expensive and prone to errors [9,10]. We introduce an alternative perspective offered by Clark [11], who argues that collaboration does not by default rely on active perspective modelling, but rather on what he refers to as ‘Common Ground’ (CG). Active modelling, then, is needed only in cases that deviate from the default created by CG [12].

We use this notion of CG to study how it can support agents in collaboration towards a joint goal, in spite of their initially different ‘views’ of a model world. We do this by implementing a version of the card game known as ‘The Game’, in which agents need to jointly play as many cards as possible, in counting order. Our agents differ in how they are initialised, such as in eagerness to play their own cards versus being more accommodating to their partners. This causes variance in their strategic repertoires – they initially have no ‘strategic Common Ground’. The game is set up such that the agents will need to coordinate their actions in each turn. This coordination can be achieved either through constantly modelling the other’s view on the game state, requiring a form of ToM, or through achieving CG on various aspects. We hypothesise that (I) accounting for the other’s perspective using a form of ToM increases performance of the agents on a collaborative task and that (II) establishing CG will retain performance, while decreasing the need for agents to actively model the other’s perspective.

We start by providing a brief background on CG and ToM, including previous work that formalised and modelled these phenomena. We then detail our computational agent-based model, and give an overview of our experimental setup. After this, we explain ‘The Game’ and how its implementation affected our model considerations. This is followed by a report and explanation of the results of our experiment, focusing on agent performance and how this is affected by ToM and CG. We discuss the implications for interaction between agents and humans, and conclude by suggesting a future extension of our model to study the benefits of CG in hybrid human-agent settings.

2. Background

2.1. Defining Common Ground and Theory of Mind

We define Common Ground (CG) as information two or more individuals have in common about a given scenario, and access to the knowledge that they both have this information [11,13], i.e., ‘reflexive shared knowledge’ (CGφ : φ. Kφφ, Kgφ, ∀p,q(Kp,qφ)). Part of CG roots in humans having the same physical structure: They have the same biology and senses as any friend or stranger. They know that they are all embodied humans, who need food and sleep to survive, and have hopes, desires, and intentions that drive them. CG is further shaped through life experiences in a process referred to as grounding [14,15]. If two individuals grow up in the same country, they share knowledge about its history, its stories, its people, and its conventions. Moreover, if two individuals share an interest, they can rely on their knowledge about that interest to model each other. Even in completely different parts of the world, two people with a love for musical theater are both familiar with recent Broadway musicals, and once they learn about the other’s love, they can immediately assume that knowledge of each other.

The concept of Theory of Mind was popularised by Premack and Woodruff when they researched whether chimpanzees have similar perspective-taking capabilities as humans.
Research in the ensuing decades showed that ToM is a complex phenomenon that emerged over millions of years of evolutionary time [17,18] and takes until at least late adolescence to fully develop in an individual’s lifetime [19,20,21,22]. The most well-known evaluation method is the ‘Sally-Anne’ task, in which pre-school children are asked to evaluate the mental model of a fictional character who had different or partial access to information in a short story [23]. As humans grow up, their ToM is greatly shaped by their culture [24,25], which influences how they interpret an event [26,27] and whether they may view someone’s actions as egocentric or collective-oriented [28,29,30]. ToM is also influenced by receptiveness to cultural differences [31,32]. In this sense, one’s CG informs their use of ToM, which should factor into models involving ToM and the establishment of CG.

There is debate over the question whether grounding relies on iteratively modelling the other’s perspective, which would imply that CG itself relies on higher-order ToM [33]. In Clark’s view, which we adopt here, this is not the case. Consider the following: People living in the same country know on which side of the road to drive and its traffic rules, and assume this of each other, making it safe to engage in traffic. Here, actions are successfully coordinated with little active thought. CG comes so naturally that it goes unnoticed – until there is a hitch. When someone suddenly starts to drive against the direction of the traffic, potential mismatches in CG come to the surface as people start to make sense of the situation: “What is this person doing? Do they not know on which side of the road one needs to drive here? Do they not want themselves and others to be safe?” The illusion is shattered – not because an active model has been violated, but because a passive assumption has failed. CG does not result from iteratively modelling others’ perspectives, but forms the basis on which such modelling can take place if needed [12].

2.2. Modelling Theory of Mind and Common Ground

Previous research has shown that agent models can be greatly beneficial to expressing ToM in a shared setting, be it competitive, collaborative, or a mix thereof [34,35,36,37,38]. We ground our model in ‘Simulation-ToM’ [39]: Agents implicitly predict their partner’s perspective in their own behaviour, rather than explicitly representing it in their reasoning (‘Theory-ToM’). Work on the ‘Tacit Communication Game’ has shown that collaboration can be achieved through non-verbal, game action, communication [42]. Our setting follows the same idea: Agents do not communicate beyond playing their cards. We are not aware of any existing models in which agents attempt to form CG by our definition.

3. Methods and Model

We have developed a simulation in which agents play a collaborative counting game by observing and adjusting their behaviour based on their fellow player’s moves. Agents implicitly assume that their partner is modelling them in response (inspired by Simulation-
ToM [39]). They are initialised to represent different views on the world – they can be more egocentric, and vary in how receptive they are towards adjusting in the direction of agents who act differently than they do. We use an existing collaborative card game, known as ‘The Game’ [43], in which players take turns to try to play every card from their hand and the deck, in the right order. The goal is to exhaust the full set of cards. Players are not aware of the order of the cards in a shuffled deck [44] and are not allowed to share which cards they drew. This makes it an imperfect-information game.

3.1. Experimental Setting and Procedure

Each player takes turns to play at minimum two cards on any of four central ‘piles’, to either count up (from ‘1’), or down (from ‘100’). Cards that deviate from the top card by ‘10’ may always be played, regardless of counting direction (an example setup can be found in Figure 1a). We use the base conditions balanced by the game designers, with an additional caveat that verbal communication is not allowed: Our setting consists of two players, who start with 7 cards in hand (each game uses a deck of 98 cards – ranging from 2 to 99). After playing 2 to 7 cards on any of the central piles, a player fills their hand back up to 7 and passes on their turn. It is important to coordinate actions to ensure that cards are played before it is too late (Figure 1b). This continues until at minimum one player can no longer play a card. The final score is based on the played number of cards, so players benefit from modelling the actions and strategies of their partners.

To win ‘The Game’, players establish communication protocols between each other to ensure they play every card, without knowing each other’s hand. In practice, this results in player protocols revealing their intentions and considerations: a ‘CG’-based communication strategy specific to their group. We evaluate the establishment of these protocols by communication-through-play. This allows us to study CG formation through observation, rather than active discourse about intentions.

3.2. Model Assumptions and Features

We do not simulate perfectly rational agents [45], as this would predefine a (CG) protocol and would not emulate human-realistic play [46,47]. Our models learn to score as high as

---

4If the deck is exhausted, players are allowed to play one card per turn (and they no longer draw draw cards).
5While purposefully kept out of scope here, such active discourse will be included in future research.
possible based on a few (unique) starting principles and iterated joint play. We initialised three settings, for ‘Low’, ‘Medium’ and ‘High’ skill agents, based on how well their interactions perform at the start of their gameplay. Every agent turn is observed by the other, who uses its own knowledge about the game to estimate the play-through strategy. If they see the other agent eagerly playing a lot of cards at the same time, they themselves likely start doing so as well – likewise, if the other agent plays fewer cards than the agent would expect, they are likely to act less eager as well. The agents in our model use three scalable features, based on interviews on human play in ‘The Game’.

- **Self-benefit** to playing more than the required two cards, based on proximity to the top card (‘Do I await more information, or is it better game-wise to play now?’)
- **Eagerness** to play more than two cards given the current partner (‘Do I want to get rid of my cards before I run the risk of being unable to play them? Do I give them more room to play cards?’)
- **Cooperativeness** towards setting (‘Does collaboration in its current form work?’).

When an agent plays a card, it finds a balance between the learned self-benefit to play the card \(s_{Ba}\) and the drive to play cards in a way that benefits collaboration with partner \(i\) \((eager_i)\), informed by difference \(dif_c\) between the card and the numerically closest card pile (difference ‘2’ is played more often than one of ‘7’). This results in equilibrium \(p_c\) (Eq. (1)). If the resulting collaboration improves performance, agents raise cooperation \(coop\) (and vice versa). At peak value, self-benefit represents a fully egocentric approach: play the minimum number of cards, as new information is always better for oneself – it is raised and lowered based on the score average over 10 rounds. Eagerness both helps the agent realise that it should play more cards before its partner renders them unplayable, and helps it realise that it plays cards so often that its partner’s cards become unplayable. Eagerness is normalised and adjusted with a \(coop\) in both scenarios (up for 5% more cards - down for 5% fewer cards).

\[
p_c : s_{Ba} \times totalNumCards > eager_i \times dif_c
\]  

(1)

Modelling in our model is implicit: Agents observe the actions of their partner, and slightly adjust their behaviour based on those actions, using ToM to know their partners do the same. We express this in ‘Eagerness’ \((eager_i)\). Successful collaboration entails correctly estimating when their partner will play a card in a given scenario, and when they themselves should. This does not mean that each agent has to respond exactly the same as the other, but it is crucial to align on what to do in a given situation. CG in turn comes from a state of mutual self-reflection. With the ToM implementation, each agent continuously updates its model based on observations of its partner. CG is assumed when an agent both (1) no longer notices a change in its partner’s behaviour (update stop \(eager_i\)) and (2) no longer notices a change in the game score resulting from its own behaviour (update stop \(s_{Ba}\)). With both of these conditions fulfilled, the agent ‘believes’ everyone is on the same wavelength about the current joint successful collaboration strategy.

The interactions in our model take place over the course of 100 rounds of 50 games. As each game consists of one randomly shuffled deck, some games are a lot harder to do well at than others. Using 50 games per round allows us to determine a decent average for evaluating our agents’ overall performance. Once an agent detects that CG exists, it stops updating its behaviour – resulting in it no longer using ToM to strategically align with its partner, nor adjusting its own self-benefit \(s_{Ba}\) - we call this CG-established (CG-est).
4. Results

Running our simulation shows that agents are able to solve the task by observing the game states and adapting to each other’s playstyles. This leads to them obtaining scores in the range of 87 - 90 (Figure 2a). As discussed, we have run the simulation under various settings, resulting in a ‘Low’, ‘Medium’ and ‘High’ skill onset. Agents stabilise towards successful interaction, despite not communicating beyond playing the cards. We observe that post-convergence strategies continue to be effective despite every game differing. The agents themselves do not ‘become’ the same: A successful solution strategy can, f.e., involve one agent acting more eagerly to account for its partner’s actions than its partner acts towards them. We highlight our results in light of our hypotheses below.

**Hypothesis 1: Accounting for the other’s perspective using a form of ToM increases performance of the agents on a collaborative task**

We have compared multiple settings of initialised agents and their performance at the task. The agents always adapt their playstyle to account for their agent-partner successfully, and informed changes in eagerness (eageri) increase the score (Figure 2a). Doing this in harmony with their partner results in collaboration: If an agent notices themselves becoming too eager (i.e., starts playing too many cards compared to their partner), they actively become less eager, to give their partner space. While the initial scores are generally low, the gameplay after learning to successfully collaborate results in 87 - 90 points. This is a significant improvement over an egocentric, self-benefit heavy, approach to the problem, where agents default to a strategy involving only self-centered play. A paired t-test comparison for 30 rounds using a collaboration strategy and 30 rounds using an egocentric strategy (Figure 2b) yields strongly significant results at $p < 0.0001$ ($t = 14.9265, df = 29, sed = 0.174$).

**Figure 2.** Figure (a) depicts the score increases over rounds for collaborative agent-agent play; Figure (b) displays the results of egocentric vs collaborative play.
Hypothesis II: Common Ground will retain model performance, allowing the agents to decrease their use of ToM. Our models show that the agents can agree on a joint strategy, once they (1) observe that the behaviour of their agent-partner no longer changes, and (2) decide that their individual behaviour no longer yields a higher score. This results in them ‘locking’ their playstyle. Afterwards, the scores remain stable, despite every game being randomly shuffled (Figure 2a): After establishing CG, agents no longer update their behaviour by actively modelling their opponent after every round, while the score does not decline. We can accept Hypothesis II under the condition that the CG is genuine: While the agents in Figure 2a have all converged to an equilibrium, agents who falsely assume CG will eventually decline into an egocentric strategy, losing model performance (this happens when CG-est does not overlap between agents). We elaborate on this in the Discussion.

5. Discussion

Our model shows that if there is sufficient CG, the agents no longer need to actively anticipate the playstyle of their partner (without losing performance). Computationally, the CG becomes a shortcut to skip the reasoning (update) steps that the agents would otherwise have to perform. This helps with resolving our aforementioned notion that actively modelling others is costly [9,48,10]. Forming CG may be a means of decreasing cognitive workload in human-human and hybrid human-machine interaction as well.

Establishing CG does occasionally fail: If agents only think they have found a solution strategy that works for both of them, the collaboration approach will eventually fall back into, or develop towards, egocentrism. This seems realistic, given that even rigid CG can fail (‘Which side of the road to drive on’, Background section). As agents cannot communicate about each other’s intentions when an ‘uncharacteristically’ complex situation arises, establishing CG without communication is quite fragile. Such fragility especially occurs when the agents perform well in one round, and then perform badly at a high number of games in the next round. The only method for an agent to break the slow adoption of egocentric strategies is to actively start modelling their partner again.

This drift to egocentrism is illustrated in Table 1. If both agents have established the two model conditions for CG at the same time, the collaboration remains stable. The more accurately CG was pinpointed, the longer the collaboration strategy stays in use. This is, however, not guaranteed: e.g., Agent a can falsely assume CG, and stops modelling agent b. Agent b continues to model agent a, and either fixes a balance that misaligns with agent a, or never ‘finds’ CG at all. This reinforces the importance of ‘grounding’ whether there actually is CG — which does rely on ToM.

### Table 1. Post-training belief $B_i$ of Agenti, in CG with Agentj (CGij) – and whether that is actually the case.

<table>
<thead>
<tr>
<th>$B_i(\neg CG_{ij})$, $\neg CG_{ij}$</th>
<th>$B_i(CG_{ij})$, $CG_{ij}$</th>
<th>$B_i(\neg CG_{ij})$, $CG_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Successful collaboration</td>
<td>Successful collaboration</td>
<td>Slow decline towards solo-strategy</td>
</tr>
<tr>
<td>Slow decline towards solo-strategy</td>
<td>Rapid decline towards solo-strategy</td>
<td></td>
</tr>
</tbody>
</table>

---

6 This almost happens for the ‘High Skill (Onset)’ (blue) in Figure 2a - rounds 9 to 11. CG is not falsely established here, because agents are still learning from each other (through $e_{gen}$). If one of two agents had concluded that its partner was no longer changing its behaviour as well, the equilibrium would have collapsed.

7 Close pinpoint: assuming CG when the partner was about to ‘lock’ its behaviour, but had not done so yet.
CG in our definition ultimately results from a collaboratively agent-written script [49]. This can result in imperfect solutions if the situation is more complex than the agents have managed to perceive. As the agents have stopped modelling each other, they will never realise that there was a better possible outcome. This fits with current literature that shows ToM to be mostly useful in more complex scenarios [8]. The CG shortcut results in a successful but suboptimal outcome that can only be overcome with cognitively complex and demanding reasoning. We hypothesise that giving our agents the ability to explicitly communicate about a strategy works as a cost-effective solution to resolve this issue: If they signal that new experience has taught them there may be a better long-term strategy, they can break the cycle by agreeing to both adapt. We leave this aspect to be explored in our future work.

6. Conclusion and Follow-up

We have modelled a counting task known as ‘The Game’ to study how agents that differ in their initial view on the model world, can reach a successful agent-agent collaboration, by taking the other’s perspective into account using a form of ToM. We highlight that an implicit mutual reflection on the success of this collaboration allows for the solution strategy to be ‘fixed’ into a joint strategy – finding Common Ground, which decreases the need to actively model one’s partner. This is an important step for formalising the relationship between CG and ToM. Understanding this relationship will in turn help reasoning about alignment in both agent-agent and agent-human collaborative environments.

The next step is modelling a more nuanced representation of differences between agents, both on a fundamental (architecture) and acquired (nurture) level. We wish to see whether our hypotheses still hold if we ask our agents to collaborate with humans of different backgrounds. Additionally, our agents have a shared goal, but there is no explicit shared intent [50]: Each agent figures out what their partner is doing, but this happens implicitly. To speak about true collaboration, the agents have to actively state their intentions and discuss before one of them takes an action that affects the interaction, instead of responding based on observations only. This once again seems to indicate that we need to introduce explicit communication to further develop the impact of our model.

Lastly, we wish to experiment with the formation of CG in other scenarios. This includes other game settings, but also research tasks more grounded in real-world practice. ‘The Game’ is useful to study the formation of CG, but is limited in its action space: CG there mostly concerns how many cards to play in any given situation. Social dilemmas are more complex than optimising strategic counting, and we wish to show that our models are similarly capable of handling such situations. One might consider resource allocation, which is heavily influenced by ToM even in early infancy under both explicit [51] and hidden conditions [52]. Additional avenues include teaching-related scenarios: Reaching CG in a teaching scenario through explicit reinforcements in one’s behaviour [53], instructions about conformity [54], or peer-to-peer teaching [55]. In such ways, we wish to contribute to a society in which ever-advancing technologies stay aligned with humans.

Acknowledgments This research is part of the Hybrid Intelligence gravitation programme – number 024.004.022, financed by the Netherlands Organisation for Scientific Research (NWO).
References

Building a Stronger Case: Combining Evidence and Law in Scenario-Based Bayesian Networks

Ludi van Leeuwen a,1, Rineke Verbrugge a, Bart Verheij a, Silja Renooij b

a Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen
b Department of Information and Computing Sciences, Utrecht University

Abstract. Existing approaches to modelling legal cases in Bayesian networks focus either on correctly representing an empirical probabilistic model of evidence traces, or on modeling alternative scenarios that can explain what happened in a case. However, neither approach legally interprets, or qualifies, aspects of a scenario as a normative legal fact. Hence, the fact that a Bayesian network representing a scenario assigns a high posterior probability to a certain victim having been killed by a certain suspect, does not imply that that suspect is guilty of murder in the legal sense, because the events in the scenario cannot be qualified as legal facts.

This paper proposes an architecture for concrete legal fact idioms that qualify events in a narrative Bayesian network. This bridges the gap between the real world and the normative legal world through so-called counts-as rules. By modeling the legal facts explicitly in the Bayesian network, we can show whether a narrative completes one or more legal fact idioms. This is demonstrated using a case study. The proposed architecture may help judges and lawyers decide on which narratives they should investigate further and which narratives are stronger than others with regard to both the evidence and the legal facts.

Keywords. Bayesian networks, Legal Modelling, Decision Support, Narrative Legal Bayesian networks

1. Introduction

Judges have to combine information from different sources and consider alternative explanations in order to come to a conclusion about a criminal case. They also have to qualify the (possible) events that have occurred as legal facts. Probabilistic reasoning using Bayesian Networks is one way in which this can be done coherently. This paper proposes a first step towards such an approach.

Bayesian networks have been used to model different criminal cases [1,2,3]: The networks combine the events that might have happened (scenario) and the evidence in a case, altogether resulting in a probability of guilt for the defendant. However, existing approaches to constructing these networks have focused on how to represent the stories
presented by the prosecution and the defense, and not on how these stories are interpreted, or qualified, by a judge. The judge needs to be sure beyond a reasonable doubt that the defendant’s acts are criminal. A criminal act consists of criminal elements. We can consider these as legal facts, in the sense that they are facts that have to be proven to legal standards. One example of such legal facts is made explicit in the definition of murder, which is the ‘premeditated intentional killing of another person’ (we here stay close to the phrasing in Dutch law). If a scenario is not complete with regards to the legal facts, e.g., if it does not show ‘premeditation’, the defendant cannot be convicted for murder, and the suspect is acquitted, even if the scenario is otherwise highly convincing.

In this paper we propose a hybrid approach to modelling criminal cases, where the normative and explicit knowledge structure of criminal law is embedded into scenario-based Bayesian networks, to show whether a scenario is complete with regards to the legal facts it aims to prove. By making the elements of the crime explicit, we can explicitly consider their probability given the evidence. In this way it can be expressed whether the probability of a legal fact is low due to a lack of evidence, or due to the scenario being incomplete. This allows both prosecution and defense to know whether their resources should go into finding stronger evidence to support part of a narrative, or to argue for or against a qualification of a (sub)-scenario as a legal fact.

The remainder of the paper is structured as follows: In Section 2, the preliminaries on Bayesian networks and narrative and probabilistic approaches to modelling criminal cases are introduced. Section 3 introduces our hybrid Bayesian network model architecture. In Section 4, the method is illustrated using an example case study. Section 5 is a discussion of the method that compares and contrasts it with other approaches.

2. Preliminaries and State of the Art

A Bayesian network (BN) is a tuple $B = (G, Pr)$, where $G$ is a directed acyclic graph (DAG) that captures a joint probability distribution $Pr$ over random variables [4]. The independences among the variables are coded in the DAG and serve for factorizing the joint distribution into conditional distributions for each variable (or node) given its parents in $G$: $Pr(V_1, ..., V_n) = \prod_{i=1}^{n} Pr(V_i | parents(V_i))$. From a Bayesian network, any prior or posterior probability of interest over a subset of the variables can be computed.

Bayesian networks have been used to model criminal cases, or aspects thereof. Existing networks range from those modelling just the interpretation and reliability of forensic evidence [5,6,7], to models including witness or expert testimonies [8], and even entire court cases including competing alternative scenarios [3,1,9]. To aid the construction of such networks, idioms can be used, which are BN fragments that represent reoccurring sub-structures and can be used as generic building blocks [10]. For example, the evidence idiom links an observation variable to an unobservable hypothesis variable in order to provide evidence for that hypothesis.

In this paper we will illustrate our proposed architecture for a criminal case described in [11]. This case has been modelled before using a non-probabilistic, argumentation-based method [12]. The latter work also studies the connection between the factual story and the legal story using constitutive rules, or count-as rules, to qualify events in the world as legal facts. Count-as rules have been used [13] and analysed [14,15] before. To the best of our knowledge, however, the connection to legal facts has not been addressed with Bayesian networks before.
3. A Hybrid Layered Bayesian Network Architecture

We propose a generic method for building Bayesian network models of criminal cases that combine three aspects of the case by means of three layers: (I) the narrative layer, (II) the evidence layer, and (III) the legal layer. The Bayesian network $B = (G, Pr)$ as a whole combines a directed acyclic graph (DAG) $G = (V, A)$ with nodes $V$ and arcs $A$, and conditional probability tables (CPTs) for each node. The set of nodes $V = V_S \cup V_E \cup V_L$ consists of three disjoint subsets of nodes, one for each layer. More specifically, the narrative layer is a sub-Bayesian network $B_S$ of $B$ with graph $G_S = (V_S, A_S)$, where $A_S = A \cap (V_S \times V_S)$. The evidence layer consists of nodes $V_E$ and arcs $A_E = (A \cap (V_E \times V_E)) \cup (A \cap (V_S \times V_E))$, that is, it includes arcs among the nodes in $V_E$ as well as arcs going from $V_S$ to $V_E$. Finally, the legal layer consists of nodes $V_L$ and arcs $A_L = (A \cap (V_L \times V_L)) \cup (A \cap (V_S \times V_L))$, that is, arcs within the layer as well as incoming arcs from the narrative layer. We will now describe the different layers in more detail.

3.1. The Narrative Layer and the Evidence Layer

The narrative layer represents one or more scenarios in $G_S = (V_S, A_S)$. A scenario is a sequence of events that attempts to explain the evidence found in the case. A scenario can correspond to a story scheme, which is a generalized pattern of events. Alternative scenarios can be presented by the prosecution and the defense. The scenarios themselves are hypothesized and not directly observed. For example, once investigators have found a dead body, but no other evidence, one scenario could be a story scheme that concludes death by natural causes, another a scheme that concludes murder, another a scheme that concludes suicide. The specific elements of the story scheme are then converted into variables $V_S$ that represent the valuation of the events, in cases of boolean variables as being true or false. The variables are then ordered temporally by $A_S$. Finally, the CPTs are specified. The result is the narrative layer in which each node represents an element of a story scheme and the nodes $V_S$ together represent all plausible scenarios that are considered by the modeller [12,2].

The evidence layer represents the evidence that was found in the specific case. It consists of nodes $V_E$, each modelling an observable piece of evidence in the real world, and forms an unconnected subgraph $G_E$ of $G$. By means of evidence idioms, nodes $V_E$ are connected to the narrative layer to either support or attack the events that are specified in the narrative layer. All nodes $V_E$ will have at least one parent in the narrative layer. For example, if one of the events in the narrative is that the suspect and the victim were seen together at some location at some time, a node in the evidence layer could be whether the suspect and victim were seen by a witness or on camera images of that location at that time. The probabilities of the events in the narrative layer should be interpreted as subjective degrees of belief, not as frequentist probabilities, because it is unclear how we should instantiate probabilities for unique cases from frequentist statistics [16]. Hence, any probabilities chosen to define the CPT will depend on the subjective degrees of belief of the modellers or investigators. The probabilities of the events in the evidence layer can be a combination of subjective degrees of belief, as well as frequentist probabilities. For events such as DNA-traces or soil recognition, there exist statistical methods that can be used to establish empirical probabilities [5]. However, for other types of more contextual evidence, subjective degrees of belief may be necessary.
3.2. The Legal Layer

The legal layer represents through its nodes \( V_L \) the normative legal facts that need to be established in order to convict a suspect of a crime. The bridge between the normative legal facts \( V_L \) in the legal layer and the hypothetical scenarios in the narrative layer, can be made using count-as rules such as of the type defined in [15,13]. This means that for each legal fact \( V_l \in V_L \), we need to consider whether there are one or more nodes \( V_s \in V_S \) in the narrative layer that have at least one valuation such that that valuation would result in the (dis-)qualification of that event as a legal fact. The nodes \( V_s \) are then taken as parents for the node \( V_l \) that represents that legal fact. The CPT of \( V_l \) is then filled out based on the values of its parents in \( V_S \). Then \( V_l \) can serve as a definition: given that the nodes in \( V_s \) have a certain valuation, this causes the node \( V_l \) to take on a certain value: either the events in that valuation of \( V_S \) count as \( V_l = true \), or they do not. If there is no relevant node in the narrative layer, then \( V_l \) has no parents. In this case, the prior probability of \( V_l \) is set to a low value in order to simulate the presumption of innocence.

We propose to cast this general pattern for legal facts in crime definitions into a Bayesian network idiom. These idioms can help us to construct the legal layer.

3.2.1. Constructing the Legal Fact Idioms

In this section, we demonstrate the design of legal fact idioms for two examples based on the Dutch criminal code for murder (Article 289 Sr) and for complicity in murder (Art 48 Sr + 289 Sr).

We can formulate the legal facts for \( Murder \) as predicates for intentional action \( Intent(x,y) \) and premeditation \( Premed(x,y) \) and \( Killed(x,y) \) which lead to \( Murder(x,y) \), where \( x, y \) are a human suspect and a victim, as based on Article 289.\(^2\) The structure of the legal fact idiom for \( Murder \) is shown in Figure 1.

![Figure 1. Legal idiom for murder](image)

For \( x, y \), the CPT for the node \( Murder(x,y) \) in the murder idiom is defined by \( Pr(Murder(x,y) = true|v_1) = 1 \) for parent valuation \( v_1 = (Intent(x,y) = true \land Premed(x,y) = true \land Killed(x,y) = true) \), and \( Pr(Murder(x,y) = true|v) = 0 \) for all other parent valuations \( v \neq v_1 \). This corresponds to the legal rule that can be expressed as: \((Intent(x,y) \land Premed(x,y) \land Killed(x,y)) = Murder(x,y)\).

To define \( Complicity in Murder (ComMu) \), we need to represent Article 48; the suspect either purposefully helped the murderer \( (Help) \), or provided aid to the murderer \( (Prov) \) (materials, information and such), and that the victim has been murdered \( (Murder) \). We can define predicates \( Murder(z,y) \), \( Help(x,z,y) \), \( Prov(x,z,y) \), where \( x \) is the human suspect (of complicity), \( y \) is the victim, and \( z \) is the murderer; we assume in this case that this murderer cannot also be the suspect of complicity.\(^3\) The structure of this idiom is shown in Figure 2. The CPT for com-

\(^2\)Further exceptions outlined in 348/350Sv in the Dutch code of criminal procedure.

\(^3\)Exceptions again in 348/350Sv.
plicity in murder $\text{ComMu}(x,y)$, for all $x,y,z$, should correspond to the legal rule
$$(\text{Murder}(z,y) \land (\text{Help}(x,z,y) \lor \text{Prov}(x,z,y))) = \text{ComMu}(x,y,z).$$
This means that for parent valuations $v_1 = (\text{Murder}(z,y) = \text{true} \land \text{Help}(x,z,y) = \text{true} \land \text{Prov}(x,z,y) = \text{true})$, and $v_2 = (\text{Murder}(z,y) = \text{true} \land \text{Help}(x,z,y) = \text{true} \land \text{Prov}(x,z,y) = \text{false})$, and $v_3 = (\text{Murder}(z,y) = \text{true} \land \text{Help}(x,z,y) = \text{false} \land \text{Prov}(x,z,y) = \text{true})$, it is the case that $Pr(\text{ComMu}(x,y,z) = \text{true} | v_i) = 1$, $i = 1,2,3$, and $Pr = 0$ for other valuations.

4. Case Study: Murder in Wamel

In this section, we illustrate our approach to model construction for a real criminal case known as “Murder in Wamel” [11] and summarised as follows (from [12]):

There are three petty criminals: Kevin, the victim; Sander, the main witness and friend of Kevin’s; and Francis, the prime suspect and an acquaintance of Kevin’s. Kevin’s body is found near two barns in the village of Wamel. He has been shot dead. Later that day, Sander contacts the police and states that he was also at the scene of the crime, allegedly trying to escape. According to Sander’s initial, later denied, statements, Francis was also at the barns and an argument developed between Kevin and Francis (allegedly over a 5000 guilders debt that Francis owed Kevin). Francis then walked to the back of one of the barns. When Kevin followed him, there was a sudden firing of shots, after which Sander fled.

We note that Sander gave conflicting testimony: First he said that Francis shot Kevin but after some time he testified that he did not see anyone else at the barn. It was known by acquaintances of Sander that Francis had told Kevin and Sander to meet up at that barn to steal weed from a third party.

We are interested in the posterior probability that Francis is guilty of murder, as well as the posterior probability that Francis is guilty of complicity in murder, considering the evidence in the case. We model the provided scenario with events in the narrative layer, then provide evidence in the evidence layer using the evidence idiom [10]. The probabilities are assigned subjectively. The events in the narrative layer are then qualified as legal facts in the legal layer. The nodes per layer and their legal qualifications are shown in Table 1, and the resulting Bayesian network is shown in Figure 3.

We collect the relevant legal idioms: Murder and Complicity in Murder, in predicate form. The variables $x = \text{Francis}$, $y = \text{Kevin}$ are instantiated in both idioms. The CPTs of Murder and ComMu remain the same when the predicates are instantiated, as the definitions of murder and complicity in the network are not dependent on the identity of the suspect. The instantiated idioms function as propositional idioms and are specific to the identity of suspect and victim: We are only investigating whether Francis murdered, or was complicit in the murder of, Kevin.

The narrative layer is connected to the instantiated legal layer. All relevant parents in the narrative layer are qualified as one of the legal facts Premed, Killed, Intent for murder and Murder, Prov, Help for complicity. If for some legal fact no relevant nodes are found in the narrative layer, then that legal fact node remains parentless.

---

4The resulting Bayesian networks will are made available at \url{https://github.com/aludi/HHAI2024}
DebtFightFK: Francis (F) and Kevin (K) have had history with each other, and had a fight.

TMathus: Testimony of witness Mathus; Sander (S) and K had both told him about F’s plan to go to the barn.

PlanBarnF: F had a plan to lure K to a barn under the pretence of stealing weed.

PlanBarnF and FightBarn together count as Premeditation: given the situation, F had a motive to lure K to the barn.

TLocation: Testimony of S that he and K were at the barn.

TSLocation: Testimony of S that he and K were at the barn.

PlanBarnF: F had a plan to lure K to a barn under the pretence of stealing weed.

PlanBarnF and FightBarn together count as Premeditation: given the situation, F had a motive to lure K to the barn.

TSF1: S's first testimony that Francis was at the barn.

Body: The forensic report of K’s body, showing cause of death.

ShootStenGun: there are three options, either no-one killed K, or F killed him, or someone who was not F killed him.

ShootStenGun and KKilled together count as Murder(z,y) in the ComMu idiom, because if someone killed K by shooting him and it was not F, then K was murdered and F might be complicit.

TSF2: S’s second testimony that he did not see F at the barn. TSF2 is conditioned on both a narrative element and on S’s first testimony TSF1.

ShootStenGun and KKilled together count as Murder(z,y) in the ComMu idiom, because if you shoot at someone and they die, then you have killed them.

KBarn: K was at the barn at the time of the murder.

SBarn: S was at the barn at the time of the murder.

FSBarn: F was in possession of a stengun (murder weapon) before the time of the murder.

TTengun: Testimony that F attempted to buy springs that could be used to repair stenguns.

TTengun: Testimony that F attempted to buy springs that could be used to repair stenguns.

TSF1: S’s first testimony that Francis was at the barn.

TSF1 and TSF2: S’s second testimony that he was at the barn. TSF2 is conditioned on both a narrative element and on S’s first testimony TSF1.

ShootStenGun = Francis counts as Intent because shooting at someone with a stengun generally shows that you intend to kill them.

SBarn and FBarn: S and F were at the barn at the time of the murder.

FStenGun: F was in possession of a stengun before the time of the murder.

FBarn: F was at the barn at the time of the murder.

KKilled: K died due to the bullets from the stengun.

Table 1. The nodes in the Narrative, Evidence and Legal layers. All nodes are boolean except ShootStenGun, which has 3 values.

4.1. Using the Bayesian Network

The resulting BN is shown in Figure 3. In this section we discuss how the posterior probability of guilt changes under different evidence valuations.

Without any instantiated evidence, the posterior probability of Francis’ guilt is low: Pr(Murder = true) = 0.06, Pr(Complicity = true) = 0.08. We now consider the case evidence. There is combined testimony that supports that Francis had debts and lured Kevin to the barn under false pretense. Moreover, we believe that Francis has a stengun and that we found Kevin’s body. However, the case hinges on Sander’s testimony. Initially, he declared that he saw Francis at the barn (TSF1). However, he declared later that he did not see Francis at the barn (TSF2). We look at the implications of all evidence, combined with Sander’s first testimony ($e_1$), then counterfactually if he had only provided TSF2 ($e_1^*$), and the combination TSF1 and TSF2 ($e_2$).

Given $e_1$ (Figure 3), we believe that Francis is at the barn, where he and Kevin fought. Therefore, the criminal-element nodes in the legal idiom for murder, premed-
Figure 3. Combined BN with evidence as in the case that only considers Sander’s first testimony. The orange nodes are instantiated evidence. The node Murdered represents Murder(z, y). Screenshot from PyAgrum BN software used for computations.

Inference in 3.01ms

Iteration, intent and killing of Kevin by Francis all have a high probability resulting in a posterior probability of murder of \( Pr(\text{Murder} = \text{true} \mid e_1) = 0.88 \) compared to the lower posterior probability of complicity in murder of \( Pr(\text{Complicity} = \text{true} \mid e_1) = 0.12 \), which still allows a possibility that someone other than Francis shot Kevin. The reason that we do not accept complicity in murder as a verdict, is the much higher probability that Francis himself was the murderer. Counterfactually, if Sander had only given his second testimony \( (e_1^*, \text{no figure shown}) \), the resulting probabilities would be \( Pr(\text{Murder} = \text{true} \mid e_1^*) = 0.02 \) and \( Pr(\text{Complicity} = \text{true} \mid e_1^*) = 0.98 \). In this case we assume that Francis was not at the barn and did not murder Kevin, and conclude that Francis was complicit in the murder due to providing information to the true murderer about Kevin’s location.

Given \( e_2 \) (no figure shown), taking into account both Sander’s testimonies, the resulting probability of guilt for either crime is not beyond reasonable doubt: \( Pr(\text{Murder} = \text{true} \mid e_2) = 0.59 \) and \( Pr(\text{Complicity} = \text{true} \mid e_2) = 0.41 \). Due to the conflicting testimony, we are unsure about whether Francis was at the barn and about who shot Sander.

5. Discussion

We have shown that we can use a legal idiom to model qualification of legal facts using counts-as rules in a Bayesian Network. The probability of the legal facts results from
the narrative, which is supported by evidence. In the following, we compare this method with existing methods for probabilistic modelling in law and qualification of legal facts. Then, modelling choices are discussed, finishing with future research on hybrid methods.

Our proposed method explicitly models the legal facts that it aims to prove and qualifies events in the scenario as legal facts. In contrast, [2] models narratives as alternative scenarios, where the probability of the scenario node is implicitly equal to the probability of guilt. However, in this theory there are no restrictions on whether the scenario qualifies the relevant legal facts. In [3], legally relevant patterns of reasoning, such as opportunity and motive, are modelled with idioms. However, while opportunity and motive are useful tools in building plausible narratives, they are not legal facts. In both these methods there is no qualification of narrative facts as legal facts. In [12], elements of the factual story can be qualified as legal facts using qualification rules, modelled as count-as rules applied to arguments. The probability of the factual story is not represented explicitly but is instead represented using notions of internal coherence, completeness, and evidential support. In contrast, in our method, we can explicitly represent the probability of each aspect of the story, and hence of the legal facts at play, resulting in a probabilistic interpretation of evidential support.

This model is based on the case study presented in [12], with some additional evidence modelled based on [11]. As this is a proof of concept, the idioms and probabilities were not elicited from a domain expert. Further application of this method requires the non-trivial qualification from narrative to legal fact and translation from law to legal idiom, which should be done by domain experts. Additionally, in this model, it is always the case that there is a set of valuations of events specified in the narrative layer that is sufficient to count as a legal fact, hence, there is a probability of 1 in the CPT of the legal fact. For other valuations, the events do not count as legal fact and this results in probabilities of 0 in the CPT of the legal idiom. However, in actual cases there might be disagreement about whether a set of narrative events counts as a legal fact (cf the ‘problem of the penumbra’ in [15]).

The proposed method could be a first step towards a hybrid system, where Bayesian networks serve as a shared representation of the facts in a case. These networks could integrate perspectives from different experts, narratives from prosecution and defense, with calculated consequences of the assigned degrees of belief, integrating statistical reasoning from empirical forensic science. Such a model would allow different users to decide where their resources should go. For example, in Figure 3, there is a high degree of belief in the legal fact of \textit{Premeditation} (Francis, Kevin) and we are less sure about \textit{Intent} (Francis, Kevin) and \textit{Killed} (Francis, Kevin). Hence, it would be useful for a prosecutor to attempt to improve support for \textit{Intent} and \textit{Killed}, and gather more evidence to make sure that it was Francis who shot the gun and not someone else. The defense should try to weaken the support for \textit{Intent} (Francis, Kevin) and \textit{Killed} (Francis, Kevin) by finding evidence to attack \textit{TStengun}. In this case, specifying degrees of belief explicitly can help each side know where to aim for further investigation.

Acknowledgements

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl.
References


The Influence of Interdependence on Trust Calibration in Human-Machine Teams

Ruben S. VERHAGEN\textsuperscript{a,1}, Alexandra MARCU\textsuperscript{a}, Mark A. NEERINCX\textsuperscript{a, b}, and Myrthe L. TIELMAN\textsuperscript{a}

\textsuperscript{a}Delft University of Technology, Delft, The Netherlands
\textsuperscript{b}TNO, Soesterberg, The Netherlands

Abstract. In human-machine teams, the strengths and weaknesses of both team members result in dependencies, opportunities, and requirements to collaborate. Managing these interdependence relationships is crucial for teamwork, as it is argued that they facilitate accurate trust calibration. Unfortunately, empirical research on the influence of interdependence on trust calibration during human-machine teamwork is lacking. Therefore, we conducted an experiment (n=80) to study the effect of interdependence relationships (complete independence, complementary independence, optional interdependence, required interdependence) on human-machine trust calibration. Participants collaborated with a virtual agent during a simulated search and rescue task in teams characterized by one of the four interdependencies. A machine-induced trust violation was included in the task to facilitate dynamic trust calibration. Results show that the interdependence relationships during human-machine teamwork influence perceived trust calibration over time. Only in the teams with joint actions (optional and required interdependence) does perceived trust in the machine not recover to its initial pre-violated value. However, results show that the correlation between perceived trust in the machine and machine trustworthiness is strongest in these teams with joint actions, suggesting a more accurate trust calibration process. Overall, our findings provide some first evidence that interdependence relationships during human-machine teamwork influence human-machine trust calibration.

Keywords. interdependence, trust calibration, human-machine teamwork

1. Introduction

Humans and intelligent machines increasingly work together as teammates on complex tasks such as manufacturing and firefighting [1]. Machines often outperform humans concerning rapid, rational, and repetitive decision-making, whereas humans are usually better at handling uncertainty and unexpected situations [2]. These separate strengths and weaknesses of humans and machines result in different dependencies, opportunities, and requirements to collaborate [3]. The ultimate goal of human-machine teams is to harness the combination of strengths of both humans and machines to accomplish what neither can do alone [4].

\textsuperscript{1}Corresponding Author: Ruben Verhagen, r.s.verhagen@tudelft.nl
Several factors determine the success of human-machine teams, for example, effectively managing the interdependence relationships between both team members [5]. Another crucial determinant is appropriate human trust in machines, meaning that they know both the potentials and limitations of machines [6,7]. A lack of appropriate trust (i.e., over- or under-trust) is one of the main reasons for the disuse and misuse of machines. This lack can be corrected by a trust calibration process over time and repeated interactions, allowing humans to adjust their expectations of the machine’s reliability and trustworthiness [6,7,8]. During the trust calibration process, repairing trust violations caused by machine errors is more difficult than building trust initially [9,10].

It is argued that interdependence relationships between humans and machines facilitate the assessment of trustworthiness of intelligent machines and accurate trust calibration by humans [11]. However, there is a lack of empirical research on the exact influence of interdependence on trust calibration in human-machine teams. For example, how different interdependence relationships during human-machine teamwork influence the trust calibration process over time is unknown. Therefore, this study investigates how complete independence, complementary independence, optional interdependence, and required interdependence influence human-machine trust calibration. To do this, we conducted a user study where participants collaborated with a virtual agent during a simulated search and rescue task in teams characterized by one of the four interdependencies.

2. Background

2.1. Interdependence in Human-Machine Teams

*Interdependence relationships* are the complementary relationships humans and machines rely on to manage dependencies during joint activities [3,12]. Joint activities concern situations in which the actions of humans depend on those of machines (and vice versa) over a sustained sequence of actions and towards a shared goal [3]. These joint activities are characterized by required, optional, complementary, or no dependencies between humans and machines, caused by their capabilities to execute actions individually and assist each other during action execution [3].

When humans and machines can both execute actions independently while collaborating towards a shared goal, they are hardly dependent on each other. On the other hand, complementary dependencies between humans and machines exist when each can only execute their unique actions that contribute to completing the overall task. Optional dependencies stem from recognizing opportunities to be more efficient when executing actions jointly rather than independently [2,3]. Finally, required dependencies originate from both team members’ lack of knowledge, skills, abilities, and resources to competently execute an action independently, but the potential to assist each other to execute the action jointly [2,3]. This distinction between complete independence, complementary independence, optional interdependence, and required interdependence essentially forms a hierarchy in coordination, dependencies, and strength of the interdependence relationship [2,3]. As these different interdependence relationships heavily affect mutual reliance and dependencies, they play a critical role in the trust relationship between humans and machines [11].
2.2. Trust in Human-Machine Teams

An early definition of trust is believing that someone or something else will act in your best interest and accepting vulnerability to this person’s or entity’s actions [13]. So, there is a trusting party (the trustor) and a party to be trusted (the trustee) [13]. Here, trust can be considered as the trustor’s perception of the trustee’s trustworthiness [7,14]. Trust is critical in all circumstances where people are in any way dependent on other’s actions, and thus more relevant in high-risk situations [7,13]. More specifically, more trust is required when the perceived risk of relying on someone or something else is higher [13]. We believe that interdependence influences the perceived risk associated with relying on someone or something else and, thus, indirectly, how much trust is required during the relationship. For example, relying on someone who can execute actions you can not is less risky than relying on someone to execute actions jointly.

Instead of blindly trusting machines, human-machine trust must be appropriate [7]. Human-machine trust is appropriate when the human’s trust in the machine is equal to the machine’s actual trustworthiness [7,15]. This match between trust and trustworthiness involves both trusting trustworthy machines and distrustng untrustworthy machines. When appropriate trust is directly caused by information about the actual trustworthiness of the machine, this is called warranted appropriate trust [16,17]. Fostering appropriate trust is crucial as a lack of appropriate human-machine trust can cause over- or under-trust in and over- or under-reliance on machines, potentially resulting in detrimental outcomes [6,7,18,19]. Fostering appropriate trust involves a process of trust calibration that corrects for over- and under-trust over time and repeated interactions, allowing humans to adjust their expectations of the machine’s reliability and trustworthiness [6,7,8].

During the trust calibration process, human-machine trust is rarely stable but instead changes over time based on past and current interactions [27,28,29,30]. Decreases in human-machine trust resulting from machine-induced trust violations can have lasting effects and are hard to recover from [29,30]. To this end, machines can deploy several trust repair strategies to repair human trust after they damage or violate it [9,30,31,32]. The most commonly used trust repair strategies include apologies, denials, explanations, and promises [29,33,34]. The impact of these trust repair strategies on human trust has been mixed, with studies showing positive, no, or even negative effects [34,35]. Moderating factors might explain these mixed results, such as the timing of the repair strategy, violation type, and violation severity [9,34]. One general result, however, seems to be the effectiveness of machine apologies for restoring trust [9,36,37]. Adding an explanation to the apology can even amplify this effect [9,38].

Explanations are not merely a trust repair strategy but also one of the primary methods for fostering appropriate human-machine trust. They specifically aim to make intelligent machines more transparent and understandable to humans [7,20,21]. Examples include machine explanations, confidence scores, and uncertainty communication, providing information about the capabilities and limitations of machines and how and why they make decisions [22,23,24]. Prior literature has shown that these forms of machine transparency can improve appropriate trust in machines [22,23,24,25,26].

2.3. Interdependence for Trust Calibration in Human-Machine Teams

In addition to machine explanations, it is argued that interdependence relationships also play a critical role in the trust calibration process [11]. In order to do so, interdependence
relationships need to be supported by observable, predictable, and directable machines [3,11]. This means that intelligent machines should be transparent and understandable enough for humans to reasonably rely on them while also allowing humans to influence their behavior [3,11]. This way, interdependence relationships can support the active and continuous exploration of trust between humans and machines to ensure that human assessments are appropriate for achieving the best possible outcomes [11].

As both trust and interdependence relationships involve risk, reliance, and dependencies, it is unsurprising that interdependence and trust are related [12]. Johnson and Bradshaw [11] argue that interdependence relationships facilitate the assessment of the trustworthiness of the machine and accurate trust calibration required for developing warranted appropriate trust. However, interdependence relationships between humans and machines can vary in terms of coordination and dependencies, such as required or optional dependencies during joint activities [2,3]. So far, there is a lack of empirical research on how these different interdependence relationships during human-machine teamwork influence human-machine trust calibration. Our study will fill that gap by comparing how complete independence, complementary independence, optional interdependence, and required interdependence influence human-machine trust calibration.

3. Method

3.1. Design

We conducted an experiment to investigate the influence of interdependence relationships during human-machine teamwork on human-machine trust calibration. To ensure a dynamic trust calibration process, we added a trust violation caused by incorrect machine advice. The experiment had a 3x4 mixed design with time as the within-subjects independent variable and interdependence as the between-subjects independent variable. Time consisted of three conditions (pre-violation, post-violation, post-recovery) and interdependence of four conditions (complete independence, complementary independence, optional interdependence, and required interdependence). As dependent variables, we measured perceived trust and the appropriate reliance rate at each of the three time points.

3.2. Participants

We recruited 80 participants through personal contacts within the university (29 female and 51 male participants). Sixty-nine participants had an age range of 18-24 years old, seven participants of 25-34 years old, one participant of 35-44 years old, two participants of 45-54 years old, and one participant of 55-64 years old. In terms of education, two participants went to high school but did not obtain a diploma, 44 participants were high school graduates, nine participants obtained some college credit but no degree (yet), one participant obtained an Associate degree, 19 participants obtained a Bachelor’s degree, and five participants obtained a Master’s degree. Concerning gaming experience, 11 participants had no experience at all, 19 participants had a little, nine participants had a moderate amount, 22 participants had a considerable amount, and 19 participants had a lot. All participants signed an informed consent form before participating in the study, approved by the ethics committee of our institution (ID 3002). Since each participant
was assigned to one of the four interdependence conditions, it was essential to control for gender, age, education, and gaming experience between these conditions. Results showed no significant differences between interdependence conditions for any of the demographic factors gender ($\chi^2(3) = 3.62, p = 0.31$), age ($W = 1.23, p = 0.75$), education ($W = 3.94, p = 0.27$), and gaming experience ($W = 0.86, p = 0.84$). Therefore, we did not further control for these demographics during data analysis.

3.3. Hardware and Software

To run this experiment, we used a laptop and the Human-Agent Teaming Rapid Experimentation (MATRX) software, a Python package for facilitating human-agent teaming research (https://matrx-software.com/). The laptop was used to launch our two-dimensional grid world created using MATRX. All subjective measures were collected using Qualtrics, while all objective measures were automatically logged using MATRX.

3.4. Environment

We built a MATRX world consisting of 14 areas, 26 collectable objects, 12 obstacles, and one drop zone (see Figure 1 for part of the world). Furthermore, we added an autonomous virtual agent (RescueBot) and a human agent (controlled by the participants) to our world. We designed an environment in which these two agents had to collaborate during a search and rescue task. To ensure an inclusive and realistic victim representation, we created the following eight victim types making up the world’s collection goal: girl, boy, woman, man, older woman, older man, cat, and dog. In addition, we created three injury types: critical, mild, and healthy. Injury type was represented by the color of the victims, where red reflected critically injured, yellow mildly injured, and green healthy victims. Eight of the 26 victims were either mildly or critically injured and had to be delivered at the drop zone, whereas the other 18 were healthy. We also added three obstacle types in front of area entrances: boulder, tree, and stone. Finally, we added flooded water to the environment, which slowed the agents’ speed as they moved through it.

3.5. Task

The objective of the task was to find the target victims in the different areas and carry them to the drop zone. Interdependence relationships between humans and RescueBot were manipulated, resulting in four conditions characterized by unique dependencies [3]. In the complete independence condition, the human and RescueBot could execute all actions independently (i.e., remove all obstacles and rescue all victims). In the complementary independence condition, RescueBot could only remove obstacles, whereas the human could only rescue victims. The other two conditions also included joint actions. In the optional interdependence condition, the human and RescueBot could execute all actions independently and jointly. However, joint action execution was four times faster than independent action execution. In the required interdependence condition, all actions had to be executed jointly. Independently removing obstacles took four seconds for stones, eight seconds for trees, and 12 seconds for boulders. Independently rescuing victims took four seconds for mildly injured victims and eight seconds for critically injured victims. Participants had ten minutes to complete the task (i.e., drop all victims at the drop zone) and received points for each victim they rescued. Rescuing critically injured
victims added six points to the total score, while rescuing mildly injured victims added three points, resulting in a maximum possible score of 36 points. Other than points and rescue time, no other differences existed between mildly and critically injured victims.

During the task, extreme rain hit the MATRX world three times: after two, four, and six minutes. This rain lasted for ten seconds and if participants did not seek shelter in one of the areas during the rain, they would lose ten points of their score and their avatar would freeze until the rain disappeared. The extreme rain merely affected score and time; it did not affect the victims to be rescued. Before the extreme rain, RescueBot warned the participants about its severity and correspondingly recommended seeking shelter or continuing with the search and rescue task. Each message was accompanied by a ping sound and color highlights to draw attention. After the rain disappeared, RescueBot provided feedback on whether the advice was correct, and more flooded water was added to the environment. RescueBot’s first advice was correct. In contrast, RescueBot’s second

Figure 1. Experimenter view of the MATRX world used for our study.
Table 1. Overview of the advice and feedback messages provided by RescueBot during the experiment.

<table>
<thead>
<tr>
<th>Message type</th>
<th>Message content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advice T1,3</td>
<td>I have detected extreme rain arriving soon and predict it will cause new floods. I advise you to take shelter in one of the areas as soon as possible, until the rain is over.</td>
</tr>
<tr>
<td>Feedback T1,3</td>
<td>My advice was correct, that weather was extreme! If you had (not) taken shelter, you would (not) have lost mission time due to injuries and 10 points of our score.</td>
</tr>
<tr>
<td>Advice T2</td>
<td>I have detected light rain arriving soon but predict it will cause no floods. I advise you to continue searching and rescuing victims.</td>
</tr>
<tr>
<td>Feedback T2</td>
<td>My advice was wrong. The amount of rain was heavy instead of light. Because of that my flood prediction was incorrect. I am really sorry.</td>
</tr>
</tbody>
</table>

advice was incorrect, provoking a trust violation. Therefore, the following feedback message contained a trust repair message explaining what happened and expressing regret [9]. We included this element of risk to the task because risk and vulnerability are critical elements of trust [7]. RescueBot’s third recommendation was correct again. Table 1 shows all the advice and feedback messages provided by RescueBot.

3.6. Agent Types

We added two agents to the world: an autonomous rule-based virtual agent (RescueBot) and a human agent controlled by the participants using their keyboards. RescueBot always moved to the closest unsearched area during the search and rescue task. Furthermore, it tracked which areas the team had searched, which victims the team had found and where, and which victims the team had rescued. RescueBot did not execute any removing or rescuing actions autonomously. Instead, it asked the participants to decide whether to remove obstacles or rescue victims independently or jointly, accompanied by a summary of the explored areas, found victims, and rescued victims (see Figure 1). This way, RescueBot’s behavior was consistent for all interdependence conditions.

Both agents could only carry one victim at a time (either independently or jointly), detect each other within two grid cells, detect and remove obstacles or pick up victims within one grid cell, and detect walls and doors from anywhere. Both agents could also communicate using the chat box shown in Figure 1. Using buttons, participants could share their actions, perceptions, assistance requests, and answers to any questions asked by RescueBot. RescueBot added the shared information to its memory and adjusted its behavior correspondingly (e.g., by not moving to the same areas as the participants).

3.7. Measures

We used self-reporting and behavior to measure perceived trust in and demonstrated reliance on RescueBot [7]. More specifically, we subjectively measured perceived user trust in RescueBot using the 5-point Likert scale for trust in explainable artificial intelligence systems [39]. This scale consisted of eight items and measured confidence in and predictability, reliability, safety, efficiency, wariness, performance, and likeability of RescueBot. We calculated the mean of these eight items as the final perceived trust score for each of the three time points separately.

In addition, we objectively logged whether participants followed the advice given by RescueBot. Based on this data, we calculated the appropriate reliance rate on Res-
cueBot. Appropriate reliance was defined as appropriate reliance on RescueBot’s correct advice at T1 and T3 and appropriate non-reliance on RescueBot’s incorrect advice at T2. Accordingly, we calculated the appropriate reliance rate at each time point by dividing the number of appropriate (non-)reliance occurrences by the number of received recommendations so far. This way, the appropriate reliance rate was a cumulative variable.

3.8. Procedure

Participants first completed a tutorial to familiarize them with the environment, controls, and messaging system. Next, participants started the actual experiment. After one minute and 45 seconds, RescueBot warned the participants about arriving rain and whether to seek shelter. After two minutes, the rain arrived and lasted for ten seconds. When the rain disappeared, RescueBot provided feedback on whether its advice was correct. After two minutes and 20 seconds, the game paused, and participants were asked to fill out the trust questionnaire for the first time. This cycle of warning, rain, feedback, and trust questionnaire was repeated two more times with similar intervals, with the other warnings arriving at three minutes and 45 seconds and five minutes and 45 seconds. The whole study lasted about 30 minutes and was conducted offline.

4. Results

4.1. Perceived Trust and Appropriate Reliance

To investigate the effects of interdependence and time on perceived trust in RescueBot (Figure 2A), we conducted both a parametric and nonparametric mixed ANOVA. We conducted both ANOVAs because the assumption of homogeneity of variances for the parametric mixed ANOVA was slightly violated at T3. Results of the parametric mixed ANOVA showed a statistically significant interaction between interdependence and time on perceived trust ($F(6, 152) = 2.83, p < 0.025, \eta^2_G = 0.042$). Results showed that the simple main effect of interdependence on perceived trust was not significant at any of the time points. In contrast, results showed that the simple main effect of time on perceived trust was significant for complete independence ($F(2, 38) = 11.1, p < 0.001, \eta^2_G = 0.18$), complementary independence ($F(2, 38) = 9.45, p < 0.005, \eta^2_G = 0.16$), optional interdependence ($F(1.38, 26.2) = 35.6, p < 0.001, \eta^2_G = 0.37$), and required interdependence ($F(1.27, 24.2) = 35.4, p < 0.001, \eta^2_G = 0.50$). Pairwise t-test comparisons using a Bonferroni correction revealed significant differences in trust scores between all time points and for all interdependencies, except between T1 and T3 for complete independence and complementary independence (Table 2 and Table 3).

To confirm these results, we ran the nonparametric rank-based mixed ANOVA [40]. Again, results showed a statistically significant interaction between interdependence and time on perceived trust ($F(4.56) = 2.29, p < 0.05$, effect size = 0.44). These results also showed that the simple main effect of interdependence was not significant at any of the time points. Moreover, the results again showed that the simple main effect of time on perceived trust was significant for complete independence ($\chi^2(2) = 13.40, p < 0.0025$, $W = 0.36$), complementary independence ($\chi^2(2) = 14.50, p < 0.001, W = 0.34$), optional interdependence ($\chi^2(2) = 30.30, p < 0.001, W = 0.76$), and required interdepen-
Table 2. Pairwise t-test and Wilcoxon comparisons for the simple main effect of time on perceived trust for each interdependence condition. Bold values show the non-significant pairwise comparisons.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Time points</th>
<th>Δ mean</th>
<th>t</th>
<th>p</th>
<th>W</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>T1 vs. T2</td>
<td>-0.67</td>
<td>3.98</td>
<td>&lt; 0.005</td>
<td>129</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td></td>
<td>T1 vs. T3</td>
<td>-0.22</td>
<td>1.66</td>
<td>0.109</td>
<td>109</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>T2 vs. T3</td>
<td>+0.45</td>
<td>-3.46</td>
<td>&lt; 0.001</td>
<td>21</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Complementary</td>
<td>T1 vs. T2</td>
<td>-0.56</td>
<td>4.02</td>
<td>&lt; 0.005</td>
<td>172</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td></td>
<td>T1 vs. T3</td>
<td>-0.23</td>
<td>1.59</td>
<td>0.055</td>
<td>146</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>T2 vs. T3</td>
<td>+0.33</td>
<td>-3.37</td>
<td>&lt; 0.001</td>
<td>23</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Optional</td>
<td>T1 vs. T2</td>
<td>-1.06</td>
<td>6.50</td>
<td>&lt; 0.001</td>
<td>207</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>T1 vs. T3</td>
<td>-0.66</td>
<td>7.10</td>
<td>&lt; 0.001</td>
<td>208</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>T2 vs. T3</td>
<td>+0.40</td>
<td>-3.53</td>
<td>&lt; 0.001</td>
<td>16</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Required</td>
<td>T1 vs. T2</td>
<td>-1.13</td>
<td>6.35</td>
<td>&lt; 0.001</td>
<td>210</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td></td>
<td>T1 vs. T3</td>
<td>-0.41</td>
<td>4.71</td>
<td>&lt; 0.001</td>
<td>150</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td></td>
<td>T2 vs. T3</td>
<td>-0.72</td>
<td>5.65</td>
<td>&lt; 0.001</td>
<td>0</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Figure 2. Interaction plots of the effects of interdependence and time on perceived trust (A) and the appropriate reliance rate (B). Error bars represent the standard errors.

Finally, pairwise Wilcoxon comparisons using a Bonferroni correction also revealed significant differences in trust scores between all time points and for all interdependencies, except between T1 and T3 for complete independence and complementary independence (Table 2 and Table 3).

To investigate the effects of interdependence and time on the appropriate reliance rate (Figure 2B), we conducted the nonparametric mixed ANOVA because of not normally distributed data. Results showed a significant main effect of time on the appropriate reliance rate ($F(1.35) = 48.06, p < 0.001$, effect size = 1.10). Pairwise Wilcoxon comparisons using a Bonferroni correction revealed significant differences between the appropriate reliance rates at T1 and T2 ($p < 0.001$) and T2 and T3 ($p < 0.001$).
Table 3. Descriptive statistics for each combination of time and interdependence condition. M refers to the mean, MR to the mean rank, SD to the standard deviation, and AR% to the appropriate reliance rate.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Time</th>
<th>M (SD) trust</th>
<th>MR (SD) trust</th>
<th>M (SD) AR%</th>
<th>MR (SD) AR%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>T1</td>
<td>3.91 (0.50)</td>
<td>153.43 (61.72)</td>
<td>0.65 (0.49)</td>
<td>136.08 (85.88)</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>3.24 (0.68)</td>
<td>82.23 (66.69)</td>
<td>0.53 (0.38)</td>
<td>103.68 (68.37)</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>3.69 (0.63)</td>
<td>128.00 (70.54)</td>
<td>0.62 (0.31)</td>
<td>121.80 (60.06)</td>
</tr>
<tr>
<td>Complementary</td>
<td>T1</td>
<td>3.87 (0.44)</td>
<td>147.40 (55.40)</td>
<td>0.80 (0.41)</td>
<td>162.40 (72.02)</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>3.31 (0.53)</td>
<td>85.05 (51.26)</td>
<td>0.53 (0.34)</td>
<td>101.35 (62.34)</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>3.64 (0.62)</td>
<td>122.95 (66.19)</td>
<td>0.62 (0.27)</td>
<td>120.10 (55.08)</td>
</tr>
<tr>
<td>Optional</td>
<td>T1</td>
<td>4.19 (0.60)</td>
<td>178.25 (64.27)</td>
<td>0.65 (0.49)</td>
<td>136.08 (85.88)</td>
</tr>
<tr>
<td>Interdependence</td>
<td>T2</td>
<td>3.13 (0.61)</td>
<td>70.58 (55.53)</td>
<td>0.45 (0.32)</td>
<td>87.03 (55.14)</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>3.55 (0.55)</td>
<td>110.93 (59.77)</td>
<td>0.58 (0.26)</td>
<td>112.88 (53.85)</td>
</tr>
<tr>
<td>Required</td>
<td>T1</td>
<td>4.14 (0.47)</td>
<td>177.78 (54.32)</td>
<td>0.80 (0.41)</td>
<td>162.40 (72.02)</td>
</tr>
<tr>
<td>Interdependence</td>
<td>T2</td>
<td>3.01 (0.62)</td>
<td>60.60 (58.84)</td>
<td>0.45 (0.22)</td>
<td>82.38 (35.88)</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>3.73 (0.30)</td>
<td>128.83 (39.17)</td>
<td>0.62 (0.17)</td>
<td>119.85 (37.21)</td>
</tr>
</tbody>
</table>

4.2. Effects of Interdependence on Reliance and Injuries

Next, we investigated if the interaction between interdependence and time on perceived trust (Figure 2A) could be explained by differences between interdependence conditions in the number of injuries or how much they relied on RescueBot. Here, the underlying assumptions were that more reliance could result in more trust [7], and more injuries (and thus lost points) in less trust. However, the already reported nonparametric mixed ANOVA only showed a significant main effect of time on the appropriate reliance rate. Results of another nonparametric mixed ANOVA also showed a non-significant interaction effect of interdependence and time on the general reliance rate \( F(3.95) = 0.83, p = 0.51, \text{effect size} = 0.26 \), and non-significant main effect of interdependence on the general reliance rate \( F(2.96) = 1.77, p = 0.15, \text{effect size} = 0.26 \). Finally, results showed that all interdependence conditions were homogeneous concerning how often they were injured by the rain \( \chi^2 (3) = 0.21, p = 0.98 \), also at T1 \( \chi^2 (3) = 2.26, p = 0.52 \), T2 \( \chi^2 (3) = 4.80, p = 0.19 \), and T3 separately \( \chi^2 (3) = 2.35, p = 0.50 \).

4.3. Accuracy of the Trust Calibration Process

Finally, for each interdependence condition, we compared the trust calibration process over time with RescueBot’s actual trustworthiness over time, expressed in terms of its advice accuracy [7,41]. More specifically, RescueBot’s advice accuracy was 100% at T1, 50% at T2, and 67% at T3. For each interdependence condition, we ran a Spearman’s rank-order correlation to assess the relationship between perceived trust in RescueBot and advice accuracy of RescueBot. Results showed a statistically significant positive correlation between perceived trust and advice accuracy for complete independence \( \rho = 0.42, p < 0.001 \), complementary independence \( \rho = 0.40, p < 0.005 \), optional interdependence \( \rho = 0.60, p < 0.001 \), and required interdependence \( \rho = 0.69, p < 0.001 \).
5. Discussion and Conclusion

5.1. Discussion

Our results show that interdependence relationships during human-machine teamwork influence human-machine trust calibration over time (Figure 2A). Across all interdependence relationships, we observe significant post-violation trust decreases compared to pre-violated trust (T2 vs. T1) and significant post-recovery trust repairs compared to post-violated trust (T3 vs. T2). However, only in the teams with joint actions (optional and required interdependence) we observe a significant post-recovery trust decrease compared to pre-violated trust (T3 vs. T1). In other words, human-machine trust does not recover to its initial pre-violated value only in the teams with joint actions (Section 4.1). Since we do not find evidence for an influence of interdependence on reliance or the number of injuries (Section 4.2), this finding can more likely be attributed to the direct influence of interdependence relationships on human-machine trust calibration.

The results further indicate that the correlation between perceived trust in RescueBot and RescueBot’s advice accuracy is significant for all interdependence relationships but strongest for the teams with joint actions (Section 4.3). This finding supports Johnson and Bradshaw’s claim [11] that interdependence facilitates accurate trust calibration. However, it also extends the claim by showing that stronger interdependence relationships with joint actions facilitate more accurate trust calibration aligning with RescueBot’s trustworthiness. This might explain why human-machine trust does not recover to its initial pre-violated value in the teams with joint actions.

We believe that the perceived risk associated with relying on machines [13] increases with the strength of the interdependence relationship, and therefore, more trust is necessary for human-machine teams with joint actions. Prior research has shown that under such conditions of increased trust necessity, over-trust can be promising for trust calibration [7,42]. Therefore, we speculate that over-reliance on the incorrect advice at T2 resulted in a more accurate trust calibration in the teams with higher trust necessity caused by joint actions. This might also explain why the stronger interdependence relationships with joint actions facilitate more accurate trust calibration aligning with RescueBot’s trustworthiness. However, follow-up research is required to support these hypothesized relationships between interdependence, risk, (over-)reliance, and trust (necessity).

Finally, we did not find evidence of an effect of interdependence on the calibration of appropriate human-machine reliance. However, timing was an important distinction between perceived trust and the appropriate reliance rate, as perceived trust was recorded after the consequences of reliance behavior. Therefore, it made little sense to compare the calibration of appropriate reliance with RescueBot’s actual trustworthiness over time, as participants could not make an informed estimate of its accuracy at T1. All in all, our results highlight that interdependence relationships are crucial to consider carefully in human-machine teams as they can influence perceived human-machine trust calibration.

5.2. Limitations and Future Work

We identify a few limitations of our study. First, we only used three time points to reflect human-machine trust calibration over time, which is a simplified representation. Even though this representation aided in capturing some critical aspects of the calibration pro-
cess, the limited temporal scope probably did not capture all nuanced aspects of trust calibration over time. Therefore, future research could increase the temporal scope of the study, facilitating a more detailed investigation of the trust calibration process.

Furthermore, we used four distinctive interdependence relationships for our interdependence conditions. Again, this is a simplified representation of human-machine collaboration, which is often characterized by a mix of all four relationships [3,43]. However, using these four distinctive relationships allowed us to examine their unique influence on trust calibration. Even though human-machine teamwork often involves a mix of all interdependencies, our results still provide developers with crucial insights. For example, how violated trust does not recover to its initial value for teams engaged in joint actions and that these teams demonstrate a more accurate trust calibration.

We identify several directions for future work. For example, investigating the interaction between interdependence and trust repair strategy on trust calibration. We speculate that specific repair strategies work better for certain interdependencies, such as promises for relationships with joint actions and explanations for independent collaboration. Future work could test these hypotheses by extending our research environment with different trust repair strategies [29,33,34]. These results could provide valuable insights allowing machines to adapt their trust repair strategies based on interdependence.

Another suggestion for future work is studying the interaction between interdependence and violation severity on trust calibration. We speculate that more severe violations will result in higher trust decreases for teams engaged in joint actions. Future work could test these hypotheses by extending our research environment to include trust violations of different severity levels, such as machine failure during action execution and incorrect machine advice. These results could provide valuable insights for developing machines adapting to interdependence relationships to address trust calibration challenges.

5.3. Conclusion

Our study shows that interdependence relationships during human-machine teamwork influence human-machine trust calibration over time. During a simulated search and rescue task with a machine-induced trust violation, only in teams with joint actions does perceived trust in the machine not recover to its initial pre-violated value. However, our findings show that the correlation between perceived human-machine trust and machine trustworthiness is strongest in these teams with joint actions. This suggests that these stronger interdependence relationships during human-machine teamwork facilitate more accurate human-machine trust calibration. Overall, our study presents some first evidence that interdependence relationships during human-machine teamwork influence human-machine trust calibration over time. Therefore, it is crucial to consider these relationships carefully during human-machine trust calibration and to conduct follow-up research on adapting trust repair strategies to interdependence.

Acknowledgements This work is part of the research lab AI*MAN of Delft University of Technology.
References


A Conversational Agent for Structured Diary Construction Enabling Monitoring of Functioning & Well-Being

Piek VOSSENa,1, Selene BÁEZ SANTAMARÍAa and Thomas BAIERA

a Vrije Universiteit Amsterdam

ORCID ID: Piek Vossen https://orcid.org/0000-0002-6238-5941, Selene Báez Santamaría https://orcid.org/0000-0003-3808-2614, Thomas Baier https://orcid.org/0000-0002-1260-3373

Abstract. We describe a Hybrid Intelligence agent that constructs a personal diary through conversation. The diary is represented in an episodic Knowledge Graph as a timeline of events, where the communication is driven by the information need of the agent. We argue that such a structured diary provides valuable information to contextualize physical, social and mental functioning and well-being for medical research and monitoring. We provide details on the formal model and implementation and demonstrate the communication by our first baseline agent. Our code is available under the MIT license on GitHub: https://github.com/leolani/cltl-diary-parent.

Keywords. conversational agent, timeline reconstruction, health, monitoring

1. Introduction

Knowing about somebody’s well-being and functioning implies knowing about their life. Hybrid Intelligence solutions that service individual needs and goals could benefit from keeping track of personal circumstances and developments, as well as people’s future plans. Especially in the context of medical support systems, in which agents need to monitor people on a regular basis, knowledge about someone’s life is crucial for understanding their perspective and making personal decisions. In many cases, such monitoring takes place over an extended period to measure the impact of, e.g. treatment, training or coaching and to be able to detect trends or unexpected patterns that can alert caretakers. There are various benefits from automating such monitoring: it can take place continuously or on a regular basis to obtain more data points, it is less invasive, and it offers more privacy as people do not need to communicate personal information and feelings to other people directly but only abstract or generalized data can be shared.

In a medical context, automated monitoring is mainly done using sensors to detect Activities of Daily Life (ADL) [1,2,3]. Such monitoring is, however, limited to activities...
that can be detected and it lacks a deeper interpretation and perspective of the patient. We therefore propose a Hybrid Intelligence approach in which a communicative agent talks with patients on a regular basis to construct a structured diary: similar to what a friend would do. As Figure 1 illustrates, our agent uses an ontology and a reconstructed timeline to ask questions regarding a user’s personal events.

Such an agent can 1) register activities (both physical and social), including those that are not observable by technological devices, 2) get more details about these activities by asking follow-up questions, 3) get information on the patient’s perspective on these activities, and 4) get information on activities that did not happen or are planned in the near future. By having a regular friendly conversation, an agent can create a timeline for somebody’s past life and future plans. In addition to reconstructing the timeline, the agent can also directly ask for specific conditions that need to be monitored for medical purposes. In comparison to a self-written diary, an agent-driven structured diary has a more explicit and standardized representation that can better generalize over the lives of different people and can stimulate and remind people regularly to provide more information, that might be relevant but missing, on their life and details of each event.

Figure 1. Structured diary reconstruction by a conversational agent that uses a fine-tuned LLM to understand the patient input as a populated graph (A-box) given a predefined ontology (T-box). Gaps in knowledge drive communication to get more information and perspectives and to validate information. Dotted grey circles represent unknown event properties that can be addressed through questions. The timeline consists of three periods: the period before the previous conversation (green), the period since the previous conversation (red) and the future (dotted). These periods define potential gaps to be filled with events.

In this paper, we describe a conversational agent for timeline reconstruction implemented in the Leolani framework [4,5] using its capability to pro-actively drive conversations, similar to [6] who also use an ontology to drive communication but for understanding cooking recipes. We begin by describing three medical Hybrid-Intelligence use-cases in Section 2, arguing that structured diaries provide valuable contexts for the monitoring of functioning and well-being of patients and being able to generalize over patients. We proceed to define the problem in Section 3 and explain how we formalize and implement the problem in Section 4.

2In this paper, we use patient synonymous to user and fragile elders at home
2. Background and motivation

Various studies address the use of sensor technology to detect activities in the context of eHealth systems. In order to define what these activities are and to facilitate sharing of data, Woznowski et al. [1] describe a hierarchical ontology for activities of daily living (ADL). Their ontology provides a good starting point for physically perceivable activities in homes. Majumder et al [2] provide a survey of research on smart home healthcare technologies. They specifically address the use of sensors and actuators that facilitate remote monitoring of the home environment (such as temperature, humidity, and smoke in the home) as well as important physiological signs (such as heart rate, body temperature, blood pressure and blood oxygen level), and activities of the occupants. Oladinrin et al [3] investigate the perception of professionals on the use of smart home technology to improve and enhance the “ageing-in-place” of elderly residents. They conclude that the development of smart home technology for health support is complex and there are various mismatching perspective. It is essential to involve the target patients and elderly both in the design but also in the monitoring as such.

Sensor-based monitoring is limited to activities that can be detected and it lacks a deeper interpretation and perspective of the patient. Activities that cannot be observed or picked up by sensor, either because they are out of reach (e.g. going out for shopping or a walk) or because they represent socially defined activities (e.g. working or planning a holiday), will be missed. Furthermore, the patient is not actively involve to provide further details, complement (contextualizing) or correct detected activities, and to express their appreciation or perspective on the activity. Finally, the patient may have certain expectations and plan with respect to the future that is not yet observable but may be important for understanding their perspective.

Deeper and more comprehensive information from the patient can be obtained by questionnaires, as shown by the InterRAI initiative. [3] They develop questionnaires as internationally standardised instruments to obtain medical data from both caretakers and patients. Similarly, patient data acquisition and monitoring is applied within protocols for Trauma Care Pathways [7] and revolving door processes in healthcare [8]). Such protocols are however restricted in the amount of data points than can be acquired over time. Especially obtaining long-term data for patients is a challenge as it requires collecting personal information beyond the scope of the direct medical care. Furthermore, questionnaires are not tailored to personal lives and are tedious to fill in, which requires discipline from the patients. Diaries and questionnaires are one-way communication processes that do not allow further interaction to clarify, correct or elaborate.

Conversational agents may offer a number of advantages to sensor-based technology and questionnaires, especially with respect to follow-up questions, addressing errors, inconsistencies, conflicts and uncertainties and possibly other information related to medical conditions. They can communicate regularly and conveniently, producing more data points in time by actively reminding and stimulating patients. The interaction can be made more engaging and empathetic than writing diaries and filling questionnaires and more robust and standardized compared to open and free diaries kept by patients.

In the next subsections, we describe three use cases under investigation as a further motivation in which structured diaries can provide context for lifestyle monitoring, clinical treatment and research purposes.

---

3https://interrai.org
2.1. Patient recovery and functioning patterns

A 2019 UN report indicates that over 46% of individuals aged 60 years and above suffer with disabilities, thereby impinging on their ability to independently perform activities of daily life. As a corollary of the surge in chronic conditions, frailty, and disabilities among older adults, healthcare utilization has also witnessed a sharp increase, as these individuals are frequently (re)hospitalized due to their disabilities [9]. Multimorbidity, i.e., the presence of more than one chronic condition in one person, is affecting 65% of individuals aged 65-84 years and approximately 82% of those aged 85 years and above. Frail older adults with multimorbidity are prone to unplanned hospitalization by e.g., disease exacerbation, disability and falls. Hospitalizations, in turn, can further exacerbate morbidity and disability with subsequent loss of self-dependency.

Counteracting this vicious circle requires close monitoring on multiple domains of health and functioning outside hospital conditions in order to obtain insight in critical conditions in their way living. However, this requires much effort from patients and an already scarcely available workforce. A conversational agent can support this monitoring. By obtaining sufficiently rich data over time, the agent can provide more context to professionals judging the medical and mental conditions of people but also offers the opportunity to infer the functioning level directly from the patient as defined by the World Health Organization in the ICF standard: the International Classification of Functioning, Disability and Health (ICF). The agent could leverage existing medical Large Language Models (LLM) fine-tuned for ICF level classification [10,11]. Likewise, the agent can keep track of the functioning level on different ICF categories such as capability to walk, eat, concentrate, their energy or mood while conversing about their lives. Any changes in functioning over time can be signaled and produce an alert for caretakers.

2.2. Diabetes

Type 2 Diabetes Mellitus is another more specific growing health concern, characterized by chronic hyperglycemia resulting from insulin resistance and impaired insulin secretion. While it is relatively well understood that lifestyle changes are effective [12], adherence to treatment remains a challenge. Reasons for this are the chronic nature of the required behavioural changes, the time investment required, and the fact that different types of change need to be maintained simultaneously [13]. Effective lifestyle recommendations should therefore incorporate the patient’s context, preferences and values in order to maximize adherence [14].

An AI system can support patients in lifestyle changes by monitoring their ADLs and by providing coaching. Such an AI system could also support in GP consults by proposing potential lifestyle change that are likely to be adopted by the patient and become a long-lasting change. The AI system would provide this advice based on the patient recent and long term diary and thus account for their drives, barriers, current lifestyle and past attempted changes. Similar to the previous use case, the same structured diary can be used as a context for specific monitoring of conditions relevant for diabetes.

2.3. Toxicity

Cancer therapy is often associated with toxicity, especially for chemotherapies that regrettably affect dividing cells outside the tumor. For brain cancer glioblastoma, toxicity
is commonly seen after application of temozolomide, a chemotherapeutic that is part of the standard therapy. In around 13% of the patients this leads to toxicities because of blood dysfunction caused by an gradual depletion of thrombocytes (i.e. blood platelets) during therapy. Patients that are treated for cancer over a longer period of time, can suffer from varies adverse side effects of the treatment. These effects show up in their daily life after the actual treatment itself. Due to the complexity of the medication, it is very difficult to obtain sufficient data on these effects. Monitoring patients well-being and perception of life can provide valuable data at scale to evaluate the impact, especially when this is paired with the clinical treatment according to a very strict schedule. This will shed light on additional care (unplanned hospitalization) and adverse events or changes in functioning that are not directly related to the treatment.

The proposed agent can be guided by the strict predefined treatment plan and activities to provide a primary calendar which is the starting point for inquiring about any (medical) events and conditions outside the protocol. Furthermore, the agent can inquire about the perspective of the patient on these events, within and outside the protocol. This information is crucial for medical professional to learn about the impact of treatments but also to adjust treatment to personal conditions and perceptions.

3. Problem description

We frame our structured diary goal as a timeline reconstruction problem that is driven by the need of the agent to become knowledgeable about both the past and any future plans. We define three types of unknowns: 1) possible and probable events and their properties, 2) possible and probable perspectives on these events and 3) factuality of events (realis or confirmed, irrealis or not confirmed and explicitly denied [15]). Note that in our approach there is no independent way of knowing what really happened except from the direct source of information provided. In other words, what the patient confirms or claims is true by lack of other evidence and what the patient denies did not happen. In order to obtain a more advanced interpretation, the agent needs to consult other sources or make independent observations.

Figure 1 shows an example of our agent interpreting events and their properties as well as their perspective in terms of associated emotions, certainty, beliefs, expectations and denials. We define a timeline as a temporal container in which events are to be placed relative to the time of the conversation. Different periods are distinguished within this container: 1) before the previous conversation, 2) the gap between the current and previous conversation and 3) the future beyond the current conversation.

The problem is defined as to use conversation to become informed about events and their properties and put them on a timeline in corresponding periods until a certain level of density and saturation is reached. In order to address this problem, the agent needs to reason over the populated Knowledge Graph and timeline to estimate missing knowledge, find knowledge that needs certification, or project to the future as expected knowledge. Missing knowledge is defined in the T-box and A-box ontology as what should be the case, what is likely the case given previous instances (analogies, probabilities), or what is expected in the temporal containers (e.g. based on habits and hobbies). The agent needs to ask questions based on the need to reconstruct and validate the timeline, given a definition what is sufficient and reasonably obtainable from the targeted person. Therefore, the agent needs to model what is sufficient and obtainable as stopping criteria.
4. Model and Implementation

In this section we introduce how we modelled and implemented our solution to this problem, focusing on the intents that drive the agent.

4.1. Model

For modeling our problem, we use the following concepts:

- Events ($E$): occurrences of an activity grounded in time.
- Activity ($Y$), Actors ($A$), Time ($T$) and Place ($P$) as properties of events.
- Mentions ($M$): expressions that make reference to events, actors, place and time.
- Claims ($C$): statements made by a source.
- Sources ($S$): the interlocutors or a third party making claims.
- Perspectives ($V$): epistemic beliefs, sentiments, or emotions of the source towards a claim.
- Timeline ($L$): a period on a temporal ruler, used to ground events in time.
- Now ($N$): the current encounter of an agent and a human.
- Previous ($R$): the previous encounter of an agent and a human.
- History ($H$): the period on the timeline before the previous encounter ($R$).
- Gap ($G$): the period in between the previous encounter ($R$) and the current encounter ($N$).
- Future ($F$): the period after the current encounter ($N$).
- Density ($D$): proportion of events per period given a threshold of the expected events.
- Saturation ($U$): proportion of knowledge obtained for an event given the expected properties.

Representing events  We use the Simple Event Model [16] as a formal model for defining events in somebody’s life. Every instance of an $sem$:Event $e_i \in (E)$ is defined by a tuple $(Y_e, A_e, P_e, T_e)$, which captures the what, who, where and when through the relations $sem$:hasActor, $sem$:hasPlace, and $sem$:hasTime.

Representing activities  For ADL, we rely on the WHO’s International Classification of Functioning, Disability and Health (ICF4). ICF measures health and disability at both individual and population levels. In this work we use the category of Activities and participation, which is subdivided into 9 sub-chapters such as Domestic life: Preparing meals, Doing housework, or Community, social and civic life: Play, Sports, Socializing. The ADL events are included in the T-box as subclasses of $n2mu$:Activity5 while individual events are $rdf$:instanceOf of $sem$:Event and a specific ADL type. We can represent constraints on the actors, place, and time using ontology types such as people, animals, places, artifacts, materials, dates, periods, e.g. washing is done by $n2mu$:Person, using $n2mu$:Water and $n2mu$:Soap in places of type $n2mu$:Bathroom or $n2mu$:Kitchen.

Representing conversations  We use the GRaSP framework [17] to represent mentions of events in conversation. Mentions are tokens (words) within utterances from a source. A mention $m_i \in (M)$ will have a $grasp$:denotes relation with a claim $c_j \in (C)$. Each mention has also an attribution $grasp$:hasAttribution that specifies the source and their perspective on a claim, where the same claim can be made multiple times by the same or different sources representing different perspectives. As perspective values, we currently use $grasp$:polarity (categorical: reals, irrealis and denied), $grasp$:certainty (scalar, be-

4https://www.who.int/standards/classifications
/international-classification-of-functioning-disability-and-health

5The namespace $n2mu$ stands for Nice-To-Meet-You and was developed in [4]
tween 0 and 1) grasp:sentiment (scalar between -1 and 1) and grasp:emotion (categorical, 27 emotions defined in [18]). The lunch example in Figure 1 is modelled as follows:

```
Event representation as RDF triples
grasp:claim1
{lunch_01 rdf:instanceOf sem:Event, icf:Lunch .
lunch_01 sem:hasActor n2mu:Carl, n2mu:Carla .
lunch_01 sem:hasPlace n2mu:Carl home .
lunch_01 sem:hasTime 2/1-2024 .
}
```

What is not mentioned in the conversation and therefore not included in this representation is, for example, what food and drinks they had for lunch. If this is defined in the ontology as possible actors of the activity, it can be used to drive questions. The density threshold \( d_i \in (D) \) is the stop criterion to ask for more events and the saturation threshold \( u_i \in (U) \) defines the upper bound for properties (SEM relations and GRaSP perspectives) to be learned about each event. Further prioritization for asking questions is based on learned probabilities from the history \( (H) \).

4.2. Implementation

Following [4,5], our agent tracks interaction using an event-bus as temporally-grounded data through which incoming signals are stored sequentially as so-called topics. Any processing module can be connected to the event-bus as a service that is sensitive to certain input topics and pushes its interpretation as output of a new topic, which can then become the input for another module. A module becomes active whenever the conditioned input topics appear on the event-bus. The architecture provides the flexibility to build any input-output pipeline using the defined modules by defining and connecting topics.

Conversation takes place when modules interpret speech or text input signals from the user (patient) and produce output topics as system responses rendered as speech or text output signals. System responses in a simple chat system can be generated directly, as in an Eliza [19] setup or a generative LLM that is prompted, or through more complex processing using multiple modules. In our case, we use an episodic Knowledge Graph (eKG) [4] to store the interpretation of the user input because 1) we want to reason more precisely about and have control over the interpretation of the signal given a person’s life and the current conversation and 2) the reasoning defines the information need or hunger of the agent and drives the communication. Besides, an eKG does not require user-specific fine-tuning with training data and can be combined with LLMs to generate responses with lower risk of hallucination.

Using a Believe-Desire-Intent protocol [20], the platform utilizes modules that define global high-level desires. This module calls other (generic) modules that interpret incoming signals as beliefs represented in the eKG and produce intents to achieve these goals.

4.3. Conversational intent

For the agent under consideration, we define the following desires and intents:

1. **high level desires** driven by the timeline:
   
   Did what was expected/planned happen:  Did you had that lunch with your sister?
What happened since the last conversation: How was your weekend?
What is planned to happen: What are your plans for tomorrow?

2. **Mid level intents** driven by necessary/possible/probable properties in the ontology:
   - Who, what, when, where: What did you have for lunch?
   - Analogies: Did she stay long again?
   - Probabilities: Did you drink wine?

3. **Low level intents** driven by possible and likely perspectives:
   - Emotion: How was it for you?
   - Certainty: Are you sure?
   - Conflicts: But you told me before that...

The overall flow of the interaction is initiated by determining the gap \((G)\) between the now \((N)\) and the previous encounter \((R)\). During previous encounters, expectations were projected into \((G)\) that should be addressed first, after which the agent can inquire about other potential and habitual events. If \((G)\) is filled according to density and satutuation conditions \((D)\), \((U)\), the agent will inquire about the expectation and plans for the future using SPARQL queries to reconstruct the timeline. For each event, the known properties and perspectives are extracted from the eKG using further queries. Figure 4.3 gives a schematic example of multiple mentions of the same event with different source perspectives (Irrealis and Denied), and a mention of as future event on a timeline.

![Figure 2. Multiple mentions of the same event with different perspectives and a mention of a future event on a timeline.](image)

The event lunch1, represented as a set of triples claim1, is mentioned multiple times in different conversations: m1\_lunch1, m2\_lunch1, m3\_lunch1. The event itself is grounded in time on Sunday, 21:01:2024 through m2\_lunch1, which lays in time period \((G)\). In the current conversation, the patient denies the event from happening which changes the perspective from Irrealis (not known as a fact) to Denied. At the same time, the patient introduces a new event in \((F)\) with an undefined date.

Addressing events that may have happened but are not mentioned yet and therefore not introduced on the timeline, is done through open questions for a period (What did you do lately?), How was your weekend?) and by iterating further (What else did you do?) or on the basis of habitual patterns from the history (Did you go to the gym?).
These questions do not require further SPARQL queries but are driven by patterns and expectations from the past and applying density and saturation conditions only.

For each ADL event considered in the high-level processing, the agent starts a sub-loop to ask for necessary and expected relations (who, what, where, when or mid-level intents), and another sub-loop to ask about epistemic and emotional perspectives relations (low-level intents). For generating questions and statements from the agent, we use the conversational history in combination with the relevant structured data from the eKG to prompt a generative LLM (decoder). For processing responses and answers to questions from the patient, we use encoder LLMs fine-tuned for emotion detection, dialogue act classification, triple extraction.

**Density and Saturation criteria**

Events and their properties can be necessary, possible and probable. Likewise, we define the density and saturation thresholds within reasonable limits but also using the personal history. Density and saturation can be reached at all three levels of Desire and Intents, where priorities can be defined in different ways. The current implementation uses configurable thresholds for the required density of events per period and for the number of properties per event. We further use configurable thresholds to give-up after a number of attempts.

New information that is integrated into the eKG typically generates a plethora of graph patterns to respond to. From these, the agent needs to select the most effective ones to reach the density and saturation criteria. Success depends on the collaboration with the user as well as on the conversational skills of the agent (natural language understanding and generation, dialogue management). Implementation-wise, selection can be random, top-down combined with random, scripted or based on policies learned through reinforcement learning. We currently implemented a baseline with scripted interaction, proceeding as described above. Within the high-level script, a specific event is targeted randomly but this can be adapted to relevance, habitual patterns or personal preferences.

5. Conclusions and future work

In this paper, we described a conversational agent that constructs a structured diary of a person’s life and perspective. Such a structured diary provides rich and valuable contextual information for medical monitoring, coaching and research. The current baseline of the agent uses a script over three levels of desires and intents. In future work, we will further enhance the agent using reinforcement learning with learned policies for intent selection and carry out empirical testing of different components and policies. For training the system, we will generate synthetic conversational data prompting generative models with synthetic structured event data as shown here. For testing, we plan to use human-human conversations.

6. Acknowledgements

This research was funded by the Vrije Universiteit Amsterdam and the Netherlands Organisation for Scientific Research (NWO) via the Spinoza grant (SPI 63-260) awarded to Piek Vossen and the Hybrid Intelligence Centre via the Zwaartekracht grant (024.004.022).
References


Human-Aware Planning for Situational Awareness in Indoor Police Interventions

Joost VOSERS a,1, Andreas BRÄNNSTRÖM a, Erik BORGLUND b, Jonas HANSSON b and Juan Carlos NIEVES a

a Department of Computing Science, Umeå University
b Police Education Unit, Umeå University

Abstract. Indoor interventions are among the most dangerous situations police officers have to deal with, mostly due to a lack of situational awareness. This work describes a planner that determines when to provide information, implemented in DLV k. It is based on the General Tactical Explanation Model, used by Swedish police during tactical interventions. The planner is envisioned to be integrated in an augmented reality tool to enhance officers’ situational awareness.

Keywords. situational awareness, human-aware planning, augmented reality

1. Introduction

Police officers deal with many dangerous situations, with indoor emergency interventions being one of the most challenging. Situational Awareness (SA) has been identified as one of the major themes in the existing challenges [1]. SA is about obtaining a thorough understanding of the situation and environment, as well as the projection of future status, and is crucial in decision-making during high-risk situations [2,3].

One way of enhancing officers’ level of SA is by providing the proper tools. An automated planner can support officers with deciding where to go and what to do, determining what course of actions would fit the user’s goals best. For such an automated planner to be human-aware, it needs to be able to adapt to the goal of the human [4]. This is important in the case of interventions where plans and goals can change constantly.

With this in mind, the research question to be explored is: How can a human-aware planning system support police officers with interventions for indoor emergencies?

In this work, we investigate the possibility of planning a path through an indoor emergency. Providing information about the environment can help with increasing officers’ SA, assisting them in situations where this is crucial. The planner is envisioned to be integrated in an augmented reality (AR) tool to present the plans and information.

The remainder of this work is structured as follows: The related work and the applied methodology are discussed. The contributions of this work consist of a further problem understanding and a human-aware planner with implementation and evaluation. The discussion and conclusion elaborate on the results and suggest future work.

1 Corresponding Author: Joost Vossers, joge0053@student.umu.se.
2. Related Work

Previous work on logic-based planning often focuses on mobile robots, designing environments for the planner to navigate through [5]. A human-aware planner operates in a space that is populated and affected by human actors [6]. Logic-based planning with a focus on the human-aware aspect includes work on planning towards the promotion of behaviour-change, where a planner is incorporated in an AR environment [7]. Moreover, in the area of plan and goal recognition there is a body of research related to the current work [8,9,10,11,12,13]. Plan recognition as planning, originally introduced by Ramirez and Geffner [8], use planning algorithms to enable an agent to recognise the goals and plans of other agents. Empathetic Planning [9] computes solutions by considering other agents’ preferences. In Emotion-aware planning [11], trajectories are generated to transition between mental states, which can be related to the concept of planning for SA. In Active Goal Recognition [10], an agent senses and acts as part of the goal recognition process. While advancing towards its goal, the agent executes sensing and world-altering actions. This relates to the method of supporting SA in the current work. In contrast to previous works, the current work incorporates actions for information retrieval, particularly considering the tactical model, in response to deficient safety conditions.

Earlier work points out difficulties in applying new technologies in the public domain with regard to trustworthiness of both the technology towards the police and of the police towards the public [14,15]. Furthermore, Sanz-Urquijo et. al point out that deployment of AI technologies for law enforcement can come with drawbacks towards the public regarding, among others, discrimination and data safety [16].

3. Methodology

Figure 1 shows the applied methodology of an iterative design process. The first step is to gather sufficient domain knowledge on the problem through a knowledge elicitation process and a literature review. Out of the literature, we selected a model upon which to expand: the General Tactical Explanation Model (GTEM) [3]. These findings are aggregated in the knowledge engineering process. The planner is realised through a logic programming approach. The plans produced by the automated planner inform us on how to deal with the problem or indicate any gaps in knowledge requiring further research.

![Figure 1. A visualisation of the methodology of the iterative design process](image)

To better understand the problem setting, we elicited meetings with a domain expert on indoor interventions from the police education unit at Umeå University. The goal of these meetings was to develop a shared understanding of the situation and its challenges. Discussed topics include standard protocols, the main goals, and the people involved.

These meeting provided two major themes of difficulty:
• **Information transfer:** providing the available information to everyone is a challenge. Information such as which rooms have been searched can be difficult to convey within appropriate time, but is essential for understanding the situation.

• **Navigation:** from a single-user perspective, it is hard to navigate an unknown building and to properly perceive the situation and environment.

The automated planner can support here by deciding when to show the available information. This would advance the user’s understanding of their environment and the situation, making it safer and easier to navigate.

To determine when to show information, we utilise the GTEM. This model is used by police in Sweden in tactical intervention scenarios [3]. One of the main points of focus is on **human limitations**, divided into three areas:

1. **Perception:** perceive the surroundings and people in it
2. **Definition:** define the situation and see whether an attack is occurring
3. **Reaction:** react and respond to the situation within adequate time

The planner is implemented in the logic-based planning language $K$, which allows for planning under incomplete knowledge [17]. The $DLV^K$ planning system implements $K$ on top of the disjunctive logic programming system DLV. The system is able to solve problems with incomplete initial states [18]. Several example programs using the $DLV^K$ planner are available online, such as the Blocksworld domain.

In the planner, a problem is represented using a **program** which consists of:

- **fluents:** the properties of a state that are relevant to a situation
- **actions:** the actions the planner can execute to change fluents
- **always:** the rules specifying the dynamics of change in the planning domain
- **initially:** the conditions that hold at the start of a situation
- **goal:** a set of values for fluents which must hold true when the planner converges

4. **Human-Aware Planner**

This section defines the different parts of the planner. The operating space and the states in it need to be defined properly and it needs to be clear how to transition between states and what the properties of a produced plan are.

We first introduce fluents, which are a way to describe the world (i.e. the state space). They convey properties of entities in the world that are deemed relevant to the domain.

**Definition 1** The set of fluents $F = \{f_1, \ldots, f_n\}$ consists of all variables in the environment that are relevant to the state space.

Considering fluents, we define the operating space of the planner. A state in this space is characterised by a combination of the fluents’ values. State transitions are realised by taking actions to change or update the fluents.

**Definition 2** A state $s_i = \{f_1 : x_1, \ldots, f_n : x_n\}$ is a mapping from fluents to values, where each fluent $f_i$ is mapped to a value $x_i$. Each state in the state space consist of a unique combination of values for fluents.

---

2 URL: https://www.dbai.tuwien.ac.at/proj/dlv/K/
Definition 3 Given two states in the state space $S$: $s_i$ and $s_{i+1}$, with $s_i \neq s_{i+1}$, an action $a \in A$ causes a transition $(s_i, a_i, s_{i+1})$, where the current state of the planner changes from $s_i$ to $s_{i+1}$ by executing action $a_i$.

The planner then has an initial state and a goal state that are part of the state space, a set of fluents, and a set of actions. It produces a plan to navigate towards the goal state. The execution of a plan results in a trajectory (a path) from an initial state to a goal state.

Definition 4 A planner $Pl$ is a function $\langle S_I, S_G, F, A \rangle \rightarrow P$ with $S_I$ the initial state, $S_G$ the goal state, $F$ the set of fluents, $A$ the set of available actions, and $P$ a produced plan.

Definition 5 A plan $P$ is a sequence of actions $\langle a_1 \ldots a_i \rangle, i \geq 0$ which, if executed in order, provide a trajectory through the state space to reach the goal state $S_G$.

5. Implementation

A characterisation of the described system is implemented in the planning language $K$. Figure 2 visualises the use case. It contains a building with several rooms. Room C1 is the starting position and room R9 is the goal state. Room C5 contains a warning, indicating a potential dangerous situation which would require a higher level of SA before entering.

The values in each room are an indication of the obstacles of gaining perception, definition, and reaction insights there, with a higher value indicating a higher difficulty (e.g. room C5 has higher obstacle values, since there is a potential dangerous situation).

All code and map visualisations are available online.\footnote{URL: https://github.com/joostvossers/HAP-SA}

5.1. Background Knowledge

The background knowledge consists of the fixed values. This information is predetermined and depends on the use case. First, it contains the spaces and the accessibility between them. The background on the spaces consists of an id and three integers. These correspond to the three obstacle levels. All rooms have some inherent difficulty in entering, therefore all rooms have basic obstacle levels of 1. To safely enter a room with
a warning indication, the user will need to overcome more obstacles - they should only enter if their SA of the situation has been appropriately increased. Therefore, the obstacle levels for a room with a warning sign are set to 3. Since rooms C1 and R9 are the initial and goal state, they have not been given inherent obstacle values.

5.2. Planner

The implementation of the planner follows the program structure, described in Section 3.

The first part of the planner are the fluents. These are the values that can change or be changed throughout the plans. We identify the following fluents:

- The user’s current location, as a space in the environment
- The user’s indication of SA, divided into perception, definition, and reaction.

\[
\text{current\_loc}(L) \text{ requires space}(L, \_, \_, \_).
\]

\[
\text{perception\_level}(P) \text{ requires } \#\text{int}(P).
\]

\[
\text{definition\_level}(D) \text{ requires } \#\text{int}(D).
\]

\[
\text{reaction\_level}(R) \text{ requires } \#\text{int}(R).
\]

The available actions are divided into two parts. First, we need to be able to move through the environment. The remaining actions consist of actions that are used to increase the perception, definition, and reaction levels. How exactly these levels are increased depends on the specific domain and use case. In this case, the focus is on when to take which action, for which the basic ‘increase’ actions should suffice.

\[
\text{move}(L) \text{ requires space}(L, \_, \_, \_).
\]

\[
\text{increase\_perception}(X) \text{ requires } \#\text{int}(X).
\]

\[
\text{increase\_definition}(X) \text{ requires } \#\text{int}(X).
\]

\[
\text{increase\_reaction}(X) \text{ requires } \#\text{int}(X).
\]

The available actions are divided into two parts. First, we need to be able to move through the environment. The remaining actions consist of actions that are used to increase the perception, definition, and reaction levels. How exactly these levels are increased depends on the specific domain and use case. In this case, the focus is on when to take which action, for which the basic ‘increase’ actions should suffice.

\[
\text{move}(L) \text{ requires space}(L, \_, \_, \_).
\]

\[
\text{increase\_perception}(X) \text{ requires } \#\text{int}(X).
\]

\[
\text{increase\_definition}(X) \text{ requires } \#\text{int}(X).
\]

\[
\text{increase\_reaction}(X) \text{ requires } \#\text{int}(X).
\]

How the fluents change based on the execution of an action is defined in the always section of the implementation. It also specifies rules and constraints for the execution of certain actions. Together, the section consists of three parts:

- tracking the perception, definition, and reaction levels
- moving through the environment
- the inertials: fluents that need to stay consistent after the planner takes an action

The three levels are updated in the same manner. After a move to a new space, they are set to 0 because of the assumption that the user has no prior knowledge on the newly entered room. The levels can be increased one step at a time, with a minimum of 0 and a maximum of 3. As an example, the rules for updating the perception level look like this:

\[
\text{caused } \text{perception\_level}(0) \text{ after move}(L).
\]

\[
\text{caused } \text{perception\_level}(P) \text{ after move}(L), \text{ perception\_level}(P).
\]

\[
\text{caused } \text{perception\_level}(X) \text{ after increase\_perception}(X).
\]

\[
\text{caused } \text{perception\_level}(X) \text{ after increase\_perception}(X1), \text{ perception\_level}(X), \ X1=X+1.
\]

\[
\text{executable } \text{increase\_perception}(1) \text{ if perception\_level}(0).
\]
executable increase\_perception(2) if perception\_level(1).
executable increase\_perception(3) if perception\_level(2).

For the move action to be possible, the new location needs to be accessible from the current location and, most importantly, the perception, definition, and reaction levels of the user need to be high enough to safely enter the next space.

executable move(L1) if acc(L1, L2), current\_loc(L2),
space(L1, P, R, D),
perception\_level(P1), P1>=P,
reaction\_level(R1), R1>=R,
definition\_level(D1), D1>=D.

The starting state consists of the starting values for the determined fluents. In this case, the user’s start position is room C1 and the perception, definition, and reaction levels are set to 0. The goal state contains the desired values for certain fluents. The final state of a plan should include the values for these fluents for it to be considered successful. In the use case, the user should get to room R9, the room with the suspect.

6. Evaluation

To evaluate the planner, we use the environment presented in Figure 2. The planner generates plans of sequential actions through the space. If executed, these result in the goal (i.e. ending in room R9) being reached. We identify three of these possible plans and compare them with the straightforward, intuitively fastest way. The plans are:

• Fast: C1 → C5 → C6 → R9 (Figure 3)
• Plan 1: C1 → C2 → C4 → R9 (Figure 4)
• Plan 2: C1 → C2 → C3 → C6 → R9 (Figure 5)
• Plan 3: C1 → R6 → R7 → R8 → R9 (Figure 6)

Depending on the specifics of the environment, different paths can be favourable against others. We compare the paths with each other on three different scenarios:

• There is no additional warning in any of the rooms
• There is a warning for potential danger in room C5
• There is a warning for potential danger in room C6

In case of a warning on the path, the user needs to receive additional information to increase their perception, definition, and reaction levels (i.e. their SA).
The amount of steps for the four paths in the three different scenarios are shown in Figure 7. The bottom part of the bars are the move actions and the top part are the increase actions. We can see that for the scenario without additional warnings, the fastest path is one of the most optimal, which is in line with expectations. In the other scenarios, there is always a path from the planner that is favourable over the fastest path. In fact, the fastest path is never strictly preferred over all three alternatives.

![Figure 5. The trajectory of plan 2](image1)

![Figure 6. The trajectory of plan 3](image2)

**Figure 5.** The trajectory of plan 2  
**Figure 6.** The trajectory of plan 3

![Figure 7. A comparison of the amount of steps necessary to traverse the paths through the different scenarios](image3)

**Figure 7.** A comparison of the amount of steps necessary to traverse the paths through the different scenarios

### 7. Discussion

The aim of this work was to construct a planner to support police officers during indoor interventions. The foundation for the planner lies in the GTEM, which models human limitations in perception, definition, and reaction. Based on these levels and a specified environment, the automated planner provides a plan through the intervention including the necessary extra steps to keep an appropriate level of SA. The current implementation is straightforward, but versatile. It is easily adaptable to a new environment and more constraints on the system actions can be added without too much difficulty.

The planner can support police officers by deciding what kind of information is necessary at which time. The calculated plans are supposed to be part of a bigger system which incorporates them in a virtual environment. In-the-field applications of AR technologies have been suggested [19,20,21,22], as well as its applicability in training [23,24]. Police officers and experts see the possibilities and are willing to test and incorporate these tools in their routines [25,26]. Figure 8 visualises an interpretation of the AR environment on top of the planning system. Some examples of additional information are shown, such as the line on the floor and the warning on the left. This visualisation provides an intuition on the tool. An actual implementation for an AR system should include more extensive research on what information officers want to see.
With visual tools, it is important to construct ones that are *helpful* rather than *intrusive* and that do not distract from the main task. Special attention should be paid to not overstimulate the user and to keep their workload as low as possible [27].

The extra information that is presented to the user guides them while keeping in mind their SA. The planning system is human-aware in that it is based on the GTEM model, which is about dealing with human limitations in situations like this. The planner and the AR environment together can guide the user safely through the intervention.

A limitation in our implementation is in the *increase* actions. It is more intuitive to define real actions that elicit specific information. This would require more research into what officers want in these situations. Our implementation provides a framework for dealing with indoor emergencies and constructs plans on *when* to present information relevant to the scene. *What* information is presented is left to the specific use case.

Lastly, it is important to ensure transparency of the system. It should be clear how the planner comes to a plan and why it decides to prioritise certain actions. More transparency also increases the trust that police officers have in the system. As mentioned before, deploying new technologies in the public domain can result in undesired changes for the public, such as a decrease in privacy and data safety. Therefore, the ways in which the *increase* actions and the rest of the system are realised should be deliberately discussed with domain experts and the people affected in general.

8. Conclusion

This work presents a human-aware planner to support police officers during interventions for indoor emergencies. The planner models the environment and provides a plan for how to move through it. More importantly, it strives to heighten the user’s sense of SA by determining at which moment to show extra information about the situation and environment. The planner is intended to be integrated in an AR tool where this extra information is virtually provided to the police officer.

Future work should explore specific types of information to present. This requires a more intensive cooperation with police officers to learn about their needs and preferences for such a tool. The next step to extending this work would be the creation of a 3D environment for the envisioned AR system. This would allow for a more extensive evaluation with police officers to see if the planner meets their requirements.

---

4Original image from Nastuh Abootalebi, Unsplash, URL: [https://unsplash.com/@sunday_digital](https://unsplash.com/@sunday_digital)
References


Abstract. People are increasingly interacting with machines embedded with intelligent decision aids, sometimes in high-stakes environments. When a human user comes into contact with a decision-making agent for the first time, it is likely that the agent’s behaviour or decisions do not precisely align with the human user’s goals. This phenomenon, known as goal alignment, has been recognised as a critical concern for human-machine teams. Prior work has focused on the effect of automation’s behavioural properties, such as predictability and reliability, on trust in human-machine interaction scenarios. However, little is known about situations where automation’s capabilities are misaligned with humans’ expectations and its impact on trust. Even less is known about the effect of environmental factors on trust. We study the relationship between intervention behaviours and trust in a simulated navigation task where the human user collaborates with an agent with misaligned goals. We evaluate trust quantitatively using intervention frequency as a behavioural measure and qualitatively using self-reports. By advancing the understanding and measurement of trust in collaborative settings, this research contributes to the development of trustworthy and symbiotic human-AI systems.

Keywords. interventions, trust, uncertainty, human-agent interaction, goal alignment

1. Introduction

There is a notable increase in interactions between people and machines with integrated intelligent decision aids. Investors use trading agents to manage trades and make their money work smartly. Vehicles embedded with intelligent technologies process and communicate real-time information to the driver. When a human user initially comes into contact with a decision-making agent, most likely programmed by someone else with a different understanding of how the agent should function, the agent’s behaviour or decisions may not fully align with what the user wants to accomplish. This problem, known as goal (mis)alignment, has been recognised as a critical concern for human-machine teams [1]. An example of goal misalignment is when Google Maps suggests an alterna-
tive shorter route while driving along a longer but scenic route. When the human recognises the goal misalignment, for instance, by noticing the agent making recommendations that are not fully aligned with the human’s interests, human interventions occur. Considering the Google Maps-assisted driving example, such an act of intervention is when the human, preferring to experience the scenic drive, rejects the alternative shorter route. In the scope of this paper, we define interventions as a human-initiated action to alter the agent’s behaviour, specifically, rejecting the decision and suggesting an alternative to modifying the decision-making process to better align with the human’s goals. In automated driving, intervention behaviour is commonly phrased as “takeover behaviour” and is a widely researched topic [2–4]. Considering the behavioural opposite of intervention—compliance, where the human accepts the agent’s decisions—misaligned goals can have disastrous outcomes, especially in safety-critical situations. For instance, in a recent incident, tourists following GPS directions were led straight into a harbour in Kailua-Kona, Hawaii [5]. Although the drivers were unharmed, tow crews had to pull the fully submerged SUV out of the water. The “flash crash” incident in 2010 [6], where humans relied on complex autonomous systems consisting of agents responsible for trading decisions, led to large monetary losses and has significantly affected regulations for US equity markets.

Compliance is often used as a behavioural demonstration of trust in human-machine interactions [7–9]. Muir [10] laid the foundation for understanding the relationship between trust and interventions in a supervisory control setting, arguing that human interactions with automation should not be viewed as a “once and for all” activity but rather a “dynamic process” where the human decides to either intervene or leave the system running under automation. The study found that automation’s unreliability affects performance, trust, and self-confidence, impacting the human operator’s decision to switch between automatic and manual modes. Prior work has focused on the automation’s properties, such as predictability [11,12] and reliability [13,14]. However, little is known about situations where automation’s capabilities are misaligned with humans’ expectations and its impact on trust. For example, while on a trip, the driver uses a GPS-enabled system to find places of interest nearby. The system, programmed to present the closest locations first, will give the driver some options. If the driver chooses the closest attraction, it is not guaranteed that they will enjoy the attraction. If the user frequently rejects a system’s suggestions because the results are not exactly what the user wants, this can lead to a loss of trust. Even less is known about the impact of environmental factors, which we define as external events occurring outside the control of the agent and the human. Environmental factors are important to consider in systems deployed in real-life conditions. For example, an unplanned road closure may happen while finding the places of interest using the GPS-enabled navigation system. We argue that such events may further impact the trust when working with agents with misaligned goals. Stated differently, our study finds evidence for miscalibrated trust [15] in environments where the human’s and agent’s goals are misaligned, evidenced through interventions. Because miscalibrated trust leads to systems’ misuse and disuse, our study warrants further investigation into trust calibration when the human’s and the agent’s goals are misaligned. Meta-reviews on the trust in human-automation interaction literature [16, 17] highlighted the limited number of studies exploring environmental factors in trust development. In this paper, we address that gap and aim to understand how the agent’s capability to act as a decision aid and the uncertainty in the environment impact trust.
We designed the human-agent interaction use case to allow the researcher to sample trust at regular intervals during the task while carefully balancing the workload of the human user. Past studies have shown that a high workload may force a human user to rely on automation [18], creating an illusion of trustworthy behaviour. This poses a challenge to capturing trust using behavioural measures. In particular, it is recommended that for a measurement like an intervention to be feasible, the workload must be low [19]. Our interaction use case involves the human completing a navigation task travelling from point A to B with the help of an automated route planning agent. The agent finds a path to the destination and reveals only the next direction from the current position. When the human intervenes and proposes a new direction, the agent automatically recalculates a new path, including the human’s suggestion, and the task continues until the destination is reached or the budgets have run out. Route planning tasks have commonly been used to evaluate trust with respect to automation reliance behaviours in human-robot [20–22] and human-agent interactions [23, 24].

Contributions. We explore how the number and frequency of interventions as a behavioural measure of trust vary according to an agent-related factor, i.e., the agent’s capability to act as a decision aid, and an environment-related factor, i.e., the uncertainty.

We model capability in our route planning use case by assuming that the human needs can be modelled by a set of criteria (e.g., distance and time) in the route planning task. Specifically, the human wants to drive the shortest distance as quickly as possible. In our design, the agent can only optimise for one criterion (either distance or time), defined as the agent’s capability. Thus, the capability is a partial alignment of what the agent optimises for and the goal of the human. Note that this differs from the agent giving faulty advice, as is much evaluated in past works on reliability and trust in human-agent interactions. The agent’s suggestions are “correct” based on the agent’s understanding of the world model, and the agent always proposes a direction from the optimal route considering the current state.

We model uncertainty as a non-determinism in the environment, where the movement of the placement marker indicating the vehicle’s current position does not solely depend on the agent’s or the human’s choices. In the GPS-assisted driving example, our uncertainty modelling is similar to a situation where the driver turns into a by-road, a traffic officer approaches, informs about a sudden road closure and forces the driver to take a direction different from the one suggested by the agent or intended by the human. We display this visually in the interactive interface, where the moving action does not always take the placement marker in the intended direction. Note that this is different from the uncertainty arising from the agent itself, such as failures or mistakes. Given this setup, we answer the following research questions:

RQ1: What effect do the agent’s capability and the uncertainty in the environment have on the intervention frequency measured at regular intervals during the task?

RQ2: What is the relationship between intervention frequency measured during the task and the user’s confidence in agent capability and environment uncertainty?

We hypothesise that the agent’s capability and uncertainty in the environment affect intervention frequency and self-reported trust and that users’ confidence in automation is lower when there is environmental uncertainty. Further, we expect a human to intervene more and have less confidence in an agent where the consequences of goal misalignment are more severe. Rather than providing a snapshot view of trust measured at a spe-
specific point (e.g., at the end of the interaction), our study provides evidence for its variability during the task. We lay a foundation for understanding the relationship between self-reported trust and human interventions when the human and agent’s goals are misaligned and when the agent’s operating environment is uncertain. In the long term, such an understanding will lead to more effective human-agent partnerships.

2. Literature Review

The navigation task we designed aligns with the collaboration principles advanced in [25]. The agent is available throughout the task and provides direction suggestions to the human (responsiveness). Whenever the human suggests an alternative direction, the agent reruns the path planning algorithm and generates a new path that includes the direction the human suggested (joint activity). The human can freely decide to use the agent’s recommendation. Trust can lead to cooperative behaviour in human-agent collaborations [11, 26] in situations where the human trustor will be at risk if not for the cooperation.

Antecedents of Trust. We adopt the trust definition advanced by Lee and See [27] for this work. Antecedents of trust are classified as automation-related, operator-related and environment-related [16]. Automation-related covers automated machines as well as embodied and virtual agents. We present capability as automation-related and uncertainty as environment-related factors.

Automation-related. In interpersonal settings, the trustee’s ability signals trustworthy perceptions to the trustor [15]. In human-automation interaction contexts, the automation’s ability is reflected in reliability, faults, predictability, transparency and automation level, which have been shown to impact trust. Reliability has been shown to positively affect trust [13, 14, 26]. Closely related to reliability, automation faults negatively affect trust. Fault occurrence frequency [28], timing of the occurrence [29], and the magnitude [30] impact trust in different ways. Automation predictability [11] positively impacts trust. Similarly, transparency (i.e., explaining the reasoning process) [31] positively impacts trust. The level of automation, mostly evaluated in automated driving scenarios [32], impacts human trust.

The automation-related factor we introduce in this study, i.e. capability, is a construct related to the agent’s ability as defined by Mayer et al. [15]. However, we advance a nuanced take on the grounded definition of the ability, which asks whether or not the agent is capable of fulfilling its commitment [33]. Capability is derived from goal alignment definition in [34]: the degree to which the human’s goal matches the AI’s programmed goal. Specifically, the agent can only plan the route by optimising for one of the two criteria the human needs. The agent’s suggestions are “correct” by its understanding of the world model, and it always provides the best recommendation given the current state and the goal that the agent optimises; the agent is neither faulty nor unreliable. A different representation of goal misalignment is presented in [35], where they consider the alignment of reward functions between collaborating agents.

Environment-related. Hancock et al. identified team collaboration and tasking as environmental factors [17]. Team collaboration refer to in-group membership, culture, communication and shared mental models. Tasking consider task type, complexity, multi-
tasking requirement and physical environment. The study found moderate effects of environmental factors on trust development. They highlight a strong need for future empirical work to study the relationships between environmental factors and trust, citing the limited number of studies available.

Hoff and Bashir’s trust model [36] classified environmental factors as an antecedent to situational trust in human-automation interactions. They define risk and benefits of using automation as environmental factors. A similar link has been established in [37], suggesting that risk impacts trust and reliance behaviours in human-automation interactions. In a high-risk driving scenario, the participants trusted and used the GPS driving advice less [38]. A study by Hoesterey and Onnasch manipulated risk by altitude and measured trust attitude and behaviour in a decision automation task [39]. Results showed that trust attitude was not affected by risk. However, trust behaviour was higher and increased during the experiment for the automation-supported group. Conflicting results reported in these prior studies further highlight the empirical necessity to improve the understanding of the relationship between trust and environmental factors. Our study takes a step towards addressing this gap.

Adopting the definition advanced by Hoff and Bashir [36], in our study, we use the uncertainty factor to manipulate the risk in the environment. Our definition of uncertainty differs from the agent-related factor predictability in Mayer’s trust model [15]. We model uncertainty as non-determinism in the environment, where external elements prevent movement in a direction suggested by the agent or human. Uncertainty in the environment does not cause the agent to replan. The agent perceives the unexpected landing position and recommends the best direction to take from that position based on the policy it has already generated.

2.1. Measuring Trust

Trust is typically measured as behavioural measurements, self-reports, and physiological measurements. We focus our literature review on behavioural measurements and self-reports because, in this study, we use interventions as a behavioural measure and self-reported trust as a post hoc measurement.

While Muir’s study affirmed intervention as a behavioural indicator for trust (or lack thereof), a recent review on measurement of trust in automation [19] recognises that behavioural measures are capable of sampling trust at a much higher rate than self-reports, which are typically administered before and after the interaction in experimental conditions. Frequent sampling generates a more accurate measure of trust because it allows the researcher to capture “the area under the trust curve” as defined in [40]. Frequent sampling allows the researchers to capture the temporal nature of trust, yet another under-explored antecedent to trust [41]. In [34], to measure the temporal trust dynamic, trajectory epistemic network analysis is used to show the evolution of trust in human-AI conversations. We demonstrate the dynamic variation of trust, measured quantitatively with interventions using a higher rate of sampling measured periodically during the task.

Behavioural measures can be active (e.g., compliance with instructions or recommendations issued by automation, operator intervenes by taking control over from automation), passive (e.g., reliance) or engaging in risk-taking in the relationship [19]. An advantage of using a behavioural measurement is that it generates a continuous metric instead of a one-time measurement at the end of the task [37]. Our study uses interven-
tion frequency during the task as an active behavioural measurement. A model predicting the trust in automated parking features consisted of the proportion of trials in which the driver intervened in addition to operator-related and automation-related constructs [2].

Prior work has recognised the contentious relationship between behavioural measurements and trust, claiming that behavioural measurements can be influenced by factors other than trust, such as workload [42] and risk [39]. Kohn et al. recommended that experiments be designed to confirm that behavioural trust measures correlate with other trust measures, such as using validated self-reported trust measurements [19]. In [43], self-reported trust (using a validated trust in automation scale in [44]) was measured periodically during the task alongside task performance-related metrics.

3. Methods

We now describe the navigation use case, how the independent variables, capability and uncertainty were operationalised, and the experimental setup.

3.1. The Navigation Simulator

We created a web-based interactive simulator where a virtual agent assisted a human user in planning a route. A valid route took the user from a starting location (home) to the destination (hospital), passing through three locations identified as landmarks (a grocery store, a construction site and a school). Although a human could enforce their own order of visiting the landmarks, the plans generated by the algorithms enforced an implicit temporal ordering for landmark visits: starting from the grocery store to the school, to the construction site, and finally the hospital. The agent used two route planning algorithms: the \(A^*\) search algorithm with the admissible heuristic landmark-cut [45] and Q-Learning [46] to generate a route to the destination, going through all un-visited landmarks. The generated route was incrementally revealed to the user one action at a time as direction suggestions. The location map is laid on a grid. Therefore, the direction suggestions were UP, DOWN, LEFT and RIGHT. We introduced risk and vulnerability to the user by constraining the route planning task to be within a time \((t)\) and distance \((d)\) budgets. We ensured that the budgets were sufficient to complete the route planning task in pilot studies. We also simulated critical events during the route planning task. An accident occurred halfway through the trip (i.e., less than half of the distance budget was remaining) where two cells on the grid became inaccessible, and the agent had to replan as its position was pushed to an adjacent cell. The interaction scenario was presented as a cover story where the participant was asked to imagine a situation where they were driving a friend to the hospital.

We designed a 10 \(\times\) 10 grid world to keep the participants’ workload at a minimum to reduce over-relying on the agent and encourage collaborative behaviour. Every trip started from the home position (bottom left of the grid) and ended at the hospital, the final position (top right). Each cell was associated with two costs: distance and time. Distance cost was uniformly distributed (cost=1) and indicated at the top-right corner of the cell. The default time cost was 1. However, some areas have a higher time cost (cost=3), indicated in red and described as “unsafe” in the cover story. The remaining budgets, visited milestones and the agent’s suggested direction were displayed to the participant. Click-
ing the Go button indicated that the participant accepted the suggestion, thus moving the placement marker in that direction and updating the budgets. The participant could intervene by clicking the Interrupt button, at which point a web pop-up screen was displayed requesting the participant to input a new direction. Then the agent replans the route with the user’s suggestion and incrementally reveals it as before².

**Modelling Agent Capability.** We model agent capability as the partial alignment of what the agent optimises for and what the participant was required to accomplish in the route planning task. The participant was instructed to complete the route-planning task without overrunning both the time and the distance budgets, that is, travel the shortest distance while avoiding unsafe areas. However, the agent can only plan a route to travel the shortest distance or avoid unsafe areas, but not both. The agent’s capability was clearly described to the participant before the task. The distance-optimising agent aimed to take the shortest path, even if this involved passing through red cells with a higher time cost. The time-optimising agent aimed to avoid red cells as much as possible, which resulted in a lengthier route.

The task introduces Assistant Dede, which optimises for distance, and the Assistant Cece, which optimises for time. We hereafter refer to $a_d$ and $a_t$ for the distance-optimising (Dede) and time-optimising (Cece) agents, respectively. The two agents differ in the consequence of goal misalignment to the human. If the agent-human collaboration results in a distance budget overrun, more likely when working with Cece, the route planning task immediately terminates. If the time budget overruns, more likely when working with Dede, the task does not terminate; however, alert messages indicating the passenger’s deteriorating health condition (as per the cover story) will be displayed repeatedly. The goal alignment problem and the respective consequences were explained to the participant before the experiment.

**Modelling Environment Uncertainty.** We model uncertainty as non-determinism in the environment, where the moving action does not land the placement marker in the intended direction at all times; the direction of the movement itself is uncertain and does not solely rely on the agent’s choice. This takes a similar form to Open AI’s FrozenLake-v0 environment [47]. We consider two levels: with and without non-determinism. Without uncertainty, the moving action lands the marker in the expected direction. With uncertainty, we assume an 80% probability of moving in the intended direction and a 10% probability of moving in either of the directions perpendicular to that intended. These two levels are referred to here as a deterministic environment without uncertainty and a non-deterministic environment with uncertainty. We use a planner that implements the $A^*$ algorithm for the deterministic environment and reinforcement learning (RL) for the non-deterministic environment.

### 3.2. Experimental Design

We adopted a between-group mixed design described in [48] to resolve RQ1 and RQ2. The study received approval from the university’s Ethics Review Board. The participants were recruited via Prolific [49]. After completing informed consent online, participants were randomly assigned to the deterministic or non-deterministic group. We ensured that

²The source code for the web-based navigation simulator is publicly available with the Creative Commons Universal License at https://github.com/sachinisw/HHAI24-Navigation-Simulator.git
participants who took part in one uncertainty condition were excluded from the other. Before the tasks, participants were asked to read the cover story carefully. Instructions to use the web application were provided in writing and via a tutorial video. Then, the participants in each group completed the route planning task once with the distance-optimising agent and once with the time-optimising agent. The agents were presented in random order.

For each landmark visited, we collected the number of times the participant rejected the agent’s suggestion. We use the number of rejections as a measure of intervention frequency. Further, similar to the process followed in [43], the user was periodically polled at each landmark point to rate their confidence in the agent’s ability to help complete the route planning task on a 10-point Likert scale. When the trip was forcibly terminated by distance budget run out, or when the participant successfully reached the hospital, they filled out a 4-question survey commonly used to measure trust in automation and has been empirically validated developed by Muir [30].

3.3. Participants

The non-deterministic group consisted of 44 participants (52% female, 42% age 44 or above). The majority had completed either a Bachelor’s degree or school-level compulsory education (32% each). The deterministic group consisted of 47 participants (35% ages 20-25, 67% female), with a 33% being college-educated with Bachelor’s degrees.

4. Evaluation

In this section, we discuss the experiments we conducted through user studies to answer the research questions (RQs) outlined in Section 1.

4.1. RQ1: Effect of Agent Capability and Environment Uncertainty on Intervention Frequency

We use the frequency of the participants rejecting the suggestions by the navigation agent to answer RQ1. We refer to this quantity as interrupts, a measure of intervention frequency. We hereafter refer to the case of a deterministic environment with an agent optimising for distance and time as DET-ad and DET-at respectively, and for a non-deterministic environment, NONDET-ad and NONDET-at respectively.

We report both the raw (non-cumulative) number of interrupts observed for each milestone in isolation and the cumulative number of interrupts, which for each milestone is the sum of the total number of interrupts for previous milestones reached and the number of interrupts at the current milestone. When observing the cumulative number of interrupts, there appears to be a greater amount in the non-deterministic environment than that of the deterministic one. Interrupts appeared to happen initially for the deterministic environment and then stop thereafter. In contrast, more interrupts appear to occur for all milestones travelled through for the non-deterministic environment, hence the increasing trend seen in Figure 2. More interrupts were observed overall for the distance-optimising agent than the time-optimising agent, which we identify the cause being due to the distance-optimising agent passing through red blocks and the participants recognising and responding to the goal misalignment.
Figure 1. Distribution of the raw interrupt frequencies for each landmark visited. TIME-OPT and DISTANCE-OPT refer to the time-optimising (\(a_t\)) and distance-optimising (\(a_d\)) agents, respectively.

Figure 2. Distribution of the cumulative number of interrupts for each landmark visited. TIME-OPT and DISTANCE-OPT refer to the time-optimising (\(a_t\)) and distance-optimising (\(a_d\)) agents, respectively.

The median (mean) cumulative number of interrupts observed at the last landmark for the distance- and time-optimising agents, respectively, were 4 and 2 (4.53 and 2.71); the distributions of the cumulative number of interrupts between the groups differed significantly (Wilcoxon rank sum test\(^3\) (\(W = 67790, p < 0.01, \text{one-tailed}\)). The median cumulative number of interrupts overall in the non-deterministic and deterministic settings observed at the last landmark were 5 and 2 (5.29 and 2.41), respectively; the distributions in the two groups also differed significantly (Wilcoxon rank sum test (\(W = 60771, p < 0.01, \text{one-tailed}\)). The distribution of the raw interrupt data throughout the visited landmarks is displayed in Figure 1. The median number of raw interrupts observed overall (across all landmarks) for the distance- and time-optimising agents were 1.22 and 0.75 respectively. The median raw number of interrupts observed overall in the non-deterministic and deterministic settings were 1.38 and 0.67 respectively. The median number of raw interrupts was 1 overall for the non-deterministic setting. The medians for all other settings were zero due to the zero-skewed data.

4.2. RQ2: Relationship Between Confidence Ratings and Intervention Frequency

At each milestone, participants were repeatedly asked to rate their confidence in the agent’s ability to help complete the navigation task on a 1 (not at all) – 10 (extremely)
Likert scale. The agents’ planning algorithms imposed a strict ordering of the milestones visited: grocery first, followed by the school, followed by the construction site and finally, the hospital. This allows us to observe the temporal dynamics of interventions. We find that the Det-\(a_t\) case has consistently the highest median confidence rating, NonDet-\(a_d\) has the lowest ratings, which decrease at a rate faster than NonDet-\(a_t\) over time. That for Det-\(a_d\) increases over time to the level of Det-\(a_t\). Using a Kruskal-Wallis rank sum test, we found that there was not a statistically significant difference in confidence scores between the visited landmarks overall (\(\chi^2 = 0.204, p = 0.9\)), although there was a statistically significant difference between the number of cumulative number of interrupts between the landmarks (\(\chi^2 = 125.36, p < 0.01\)). A summary of the confidence ratings for each agent-environment setting is illustrated in Figure 3.

When looking at the confidence ratings overall, however, we are able to find that the median confidence rating for \(a_d\) and \(a_t\) respectively were 6 and 7, suggesting that participants were less confident in the distance-optimising agent; the distributions in the two groups differed significantly (Wilcoxon rank sum test (\(W = 32229, p = 0.011\), one-tailed)). The median confidence rating in the non-deterministic and deterministic settings were also 6 and 7 respectively, suggesting that participants were less confident in agents operating in the non-deterministic setting; the distributions in the two groups also differed significantly (Wilcoxon rank sum test (\(W = 26138, p < 0.01\), one-tailed)). From this, we can deduce that overall, a higher number of interventions is linked to lower confidence in an agent.

### 4.3 Posthoc Evaluation

We administered the validated questionnaire in [30] after the participants completed the navigation task to measure trust qualitatively.

The summary statistics (mean, median and standard deviation) of the four questions overall (\(O\)) and for \(a_d\) and \(a_t\) for the deterministic (\(D\)) and non-deterministic (\(N\)) settings are displayed in Table 4.3. Here we see a similar pattern in the post-hoc trust scores as we saw previously in the confidence ratings collected during the task. The median posthoc trust score for \(a_d\) and \(a_t\) respectively were 5 and 7, suggesting that participants had less trust in the distance-optimising agent; the distribution of scores between \(a_d\) and \(a_t\) differed significantly (Wilcoxon rank sum test (\(W = 3445.5, p < 0.01\), one-tailed)). The median trust score rating in the non-deterministic and deterministic setting were 4 and 7 respectively (highlighted in bold in Table 4.3, suggesting that participants had less trust in goal misaligned agents operating in the non-deterministic setting; the distribution of scores between the non-deterministic and deterministic settings differed signifi-
Table 1. Summary statistics for post-hoc questionnaire responses overall (O), for distance-optimising (ad) and time-optimising (at) agents under deterministic (D) and non-deterministic (ND) settings. For each question–setting combination, the mean, median and standard deviations are reported.

<table>
<thead>
<tr>
<th>Setting</th>
<th>Question 1</th>
<th>Question 2</th>
<th>Question 3</th>
<th>Question 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-ad</td>
<td>6.92</td>
<td>7.00</td>
<td>1.84</td>
<td>6.65</td>
</tr>
<tr>
<td>D-at</td>
<td>7.52</td>
<td>8.00</td>
<td>1.72</td>
<td>7.62</td>
</tr>
<tr>
<td>D-O</td>
<td>7.22</td>
<td>8.00</td>
<td>1.79</td>
<td>7.14</td>
</tr>
<tr>
<td>ND-ad</td>
<td>4.76</td>
<td>4.00</td>
<td>2.04</td>
<td>4.37</td>
</tr>
<tr>
<td>ND-at</td>
<td>5.27</td>
<td>5.00</td>
<td>2.09</td>
<td>5.06</td>
</tr>
<tr>
<td>ND-O</td>
<td>5.01</td>
<td>5.00</td>
<td>2.07</td>
<td>4.71</td>
</tr>
</tbody>
</table>

The quantitative results show that during the task the agent’s capability and the uncertainty in the environment had an impact on intervention frequency and confidence. Specifically, participants intervened more frequently for the distance-optimising agent, and when there was uncertainty in the environment. When confidence ratings were looked at, participants had a pattern of lower confidence in the distance-optimising agent compared to the time-optimising agent, and lower confidence when there is uncertainty in the environment compared to when there is a lack thereof. The self-reported trust, measured after the task, was higher on average in the deterministic setting for both agents than in the non-deterministic setting. The time-optimising agent was associated with higher trust scores on average compared to the distance-optimising agent, within the deterministic and non-deterministic settings. Our observations support the hypothesis that humans would intervene more, have less confidence in, and consequently have less trust in an agent with goal misalignment when the misalignment can be quickly recognised from its behaviour during the task (e.g., distance optimising agent travelling through red zones). The consequence of goal misalignment when working with the distance-optimising agent was less severe compared to the time-optimising agent, which may also explain the inclination to interrupt more. Further, the results show evidence that intervention frequency (when used as a behavioural measurement for trust) is indicative of self-reported trust, thus agreeing with Muir et al. findings [30].

Since we collected data for each milestone visited in a specific order, we were also able to observe temporal trends in the number of interventions. The changes in the number of interventions between different landmarks were statistically significant.

There appeared to be a sharper increase in the number of interrupts over time for the non-deterministic environment compared to that for the deterministic environment. In other words, participants tend to interrupt more as the task progresses in the non-deterministic environment. We note that our data is limited in the sense that we only have four temporal points for interrupts and three for confidence ratings, so we do not perform a correlation test here.
Working with agents programmed by others, who may not necessarily take the present user’s objectives in mind, leads to goal misalignment, which impacts interventions. We see that more uncertainty in the environment further exacerbates the problem, impacting confidence in the agent and trust. Combining the qualitative and quantitative data, we see a more detailed picture of trust dynamics in HAI. This supports Kohn et al. [19] recommendation to have both kinds of measurements in trust experiments.

Our results suggest that the goal misalignment in human-agent interaction scenarios negatively impacts trust. We observe that being aware of the goal misalignment at the start of the interaction did not maintain trust. This finding implies that such situations require the agent/automation to be built with mechanisms to build and foster trust continuously throughout the interaction, such as providing explanations or utilising design characteristics that signal trustworthy perceptions.

Further, Kohn et al. [19] suggest that behavioural metrics are affected by external aspects such as workload and risk, motivating the need to confirm that behavioural trust measures correlate with other trust measures. Using validated self-reported measures is one solution. We measured self-reported trust only at the end of the task. We can measure trust between the milestones and wish to perform this comparison in the future. Considering the limitations of our study, it would be ideal to have a finer-grained notion of the temporal dynamics of trust, as we only consider four milestones during the task and collect intervention frequencies and confidence scores. There may be human factors-related causes for the observed trends, which cannot be explained with our data.

6. Conclusion & Future Work

In this work, we examined what interventions reveal as a behavioural measure of trust in an agent-assisted collaborative environment. We show that the agent’s capability and the uncertainty in the environment had an impact on intervention frequency and the human’s confidence in the agent during the task. We also found that the goal misalignment between a human and an agent impacts trust. Similarly, trust decreases even further in non-deterministic settings. In real-life settings where it is infeasible for the human to avoid uncertainty or goal misalignment, the agent/automation needs to be embedded with features or capabilities that reinforce trust. Our study assumed that the human’s need are to optimise for distance and time, while the agent can optimise only for one, which remains static throughout the interaction. In the future, we aim to explore how trust evolves when the agent can adapt to human preferences. We also wish to expand the analysis into other human factors, such as self-confidence and affinity for technology, which have been shown to impact trust in automation [50].

Acknowledgements

This research was funded by the UK Research and Innovation Trustworthy Autonomous Systems Hub (EP/V026801/2). We thank Rita Borgo and Luca Viganò for their guidance in designing the study and the anonymous reviewers for their feedback.
References

[44] Jian-Yin Jian AMB, Druty CG. Foundations for an Empirically Determined Scale of Trust in Automated


Using Default Logic to Create Adaptable User Models for Behavior Support Agents

Johanna WOLFF a,1 Victor DE BOER bDirk HEYLEN aand M. Birna VAN RIEMSDIJK a
aUniversity of Twente
bVrije Universiteit Amsterdam

Abstract. Behavior support agents can assist a user in reaching their goals by suggesting suitable actions. In order for these agents to be effective, the agent’s advice should be personalized to the user’s needs and preferences. However, the way context influences the user, the internal state of the user and the user’s desired behavior are all subject to change while the agent is in use. If the agent is not able to adapt to these changes, this can lead to a misalignment between the user and the agent. By making the reasoning of the agent explicit, we can allow the user to directly interact with the agent’s user model in order to resolve possible misalignments. We propose to use ordered default logic to reason about the user model as its defeasible nature is inherently well suited to model behavior patterns and routines which may have exceptions dependent on the context. We then analyze different misalignment scenarios and describe how we can use various belief revision techniques to update the agent’s user model and resolve these misalignments.

Keywords. Default Logic, Non-Monotonic Logics, Belief Revision, Human-Machine Alignment, User Modeling, Behavior Support Agents

1. Introduction

Artificial assistants which are designed to help their users change their behavior and adopt new routines [1] are becoming increasingly popular. These agents are most effective when they are personalized to the user’s needs and preferences [2,3]. Beyond that, artificial support agents are increasingly expected to work as a team together with the human user [4]. However, the way context influences the user, the internal state of the user and the user’s desired behavior are all subject to change while the agent is in use. The agent needs to be able to adapt to these changes in order to support the user over a longer period of time [3]. This ensures not only the effectiveness of the agent but also that the user remains in control of how they use the technology [4,5].

Machine learning techniques can be used to achieve a high degree of personalization [6], but these data-driven approaches can also make it difficult to update or influence the knowledge base directly when the user changes their behavior or preferences. This is because relevant concepts are often not explicitly represented, making it unclear how to input new information and which effect this may have on the agents output [7].

1Corresponding Author: Johanna Wolff, j.d.wolff@utwente.nl.
To allow the user to interact with the reasoning of the agent directly and make necessary changes in case of a misalignment, we use knowledge-based methods to represent this explicitly. In particular, we propose a way to express the user model of a behavior support agent in ordered default logic (Section 2). As default logic uses defeasible reasoning, this allows us to draw tentative conclusions based on incomplete information about the world. The default rules are used to represent the agent’s beliefs about the user, in particular about their behavior patterns and routines (Section 3). By including a preference relation on these defaults we can represent the user’s preferences and priorities.

If the user feels like the agent is not providing optimal suggestions, this needs to be resolved by adapting the agent’s reasoning. We analyze different misalignment scenarios based on [8] and show that these can each be resolved by performing an appropriate belief revision update on the default logic (Section 4).

Example 1 Throughout this paper we will illustrate our work using an example of a simple scheduling agent which helps the user find time to exercise. By taking the user’s schedule, goals, routines and preferences into consideration, the agent attempts to find the best possible suggestion. The agent is also able to process information about certain contexts, and otherwise relies on the user to add additional information which is needed.

2. Preliminaries

Default logic was first introduced in [9] as a way to reason with beliefs which may need to be rejected when additional information is obtained. The logic is characterized by the introduction of default rules of the form

\[
\text{Prerequisite} : \text{Justification} \quad \delta
\]

\[
\text{Consequent}
\]

which express that if the prerequisite is given and there is no proof that the justification is false, then we infer the consequent.

In this paper we use the framework introduced in [10], which includes an ordering on the default rules of a theory. A theory of this ordered default logic can be translated into standard default logic, allowing for an implementation in theorem provers for standard default logic. An ordered default theory has the form \( T = (W, D, <) \) where \( W \) is a set of proposition logic formulas, \( D \) is a set of default rules and \(<\) is a strict partial order on \( D \). The sentences in \( W \) describe our, possibly incomplete, knowledge of the world, while the default rules in \( D \) allow us to derive additional information based on our beliefs. For two default rules \( \delta_1, \delta_2 \), we take \( \delta_1 < \delta_2 \) to mean that \( \delta_1 \) can only be applied after \( \delta_2 \) has either been applied or blocked. A default rule can be blocked either because the prerequisite cannot be proven or the negation of the justification has been proven.

A minimal set of sentences \( E \) which contains \( W \) and is deductively closed regarding both the default rules in \( D \) and standard logical inference is called an extension of the theory \( T \). In [10] formal requirements are given which enforce the existence of a consistent extension, the detailed proof is taken from [11]. The technical details are outside the scope of this paper, but necessary to ensure the usability of the framework.

The logic of belief revision is concerned with the formalization of different change operators on a set of beliefs. In particular, these operations should be in line with relevant
rationality postulates and result in a consistent set of beliefs. One of most dominant theories of belief revision is the AGM model. We will be using the following operators based on [13].

<table>
<thead>
<tr>
<th>Operator</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W \div \varphi )</td>
<td>Completely removes the sentence ( \varphi ) from ( W )</td>
</tr>
<tr>
<td>( W \ast \varphi )</td>
<td>Completely removes ( \neg \varphi ) from ( W ) and adds ( \varphi )</td>
</tr>
<tr>
<td>( D \div \delta )</td>
<td>Removes the default rule ( \delta ) from ( D )</td>
</tr>
<tr>
<td>( D \ast \delta )</td>
<td>Adds the default rule ( \delta ) to ( D )</td>
</tr>
<tr>
<td>( T \ast_{+} 1 \varphi )</td>
<td>Ensures that there is at least one extension of ( T ) which contains ( \varphi )</td>
</tr>
</tbody>
</table>

By using the translation of ordered default logic into standard default logic, we can apply the belief revision operators introduced here to our framework. When updating the default rules in \( D \), there also needs to be corresponding update which adjusts the ordering \( < \) accordingly. In the following we consider these changes to be included in the updates.

3. Using Ordered Default Logic for User Modeling

We propose to use ordered default logic as described in Section 2, to reason about the agent’s user model and determine which advice the agent should present to the user. The input of the agent’s reasoning process is a theory of ordered default logic \( T = (W, D, <) \) which describes the user model and the knowledge of the world, the output is a set of extensions \( E \) of this theory which determine the advice that the agent presents to the user.

The user model of our framework is based on the preference-based reasoning for BDI agent systems introduced in [14]. Specifically, the user model will include the user’s goals, the possible actions that can be taken to achieve these goals, behavior patterns and the preferences that the user has regarding these. Additionally, we include knowledge and beliefs about the world, which allow us to reason about the context that the other concept can appear in. In the following, we detail how each of these concepts can be represented in a theory of ordered default logic.

Example 2 In Table 1 we show how the agent from Example 1 can be represented using our framework. We only look at the schedule for one day, and only differentiate between morning and afternoon, abbreviated as Morn and Aftn respectively. We use the predicates \( \text{Plan}(t, a) \) to express that an action \( a \) is scheduled at time \( t \) and \( \text{Friend}(t) \) to express that a friend is available at time \( t \). Together, this information in Table 1 forms the theory \( T = (W, D, <) \) with the extensions \( E_1, E_2 \) and \( E_3 \). For readability we have only listed the sentences which are relevant for the agent’s advice.

Knowledge and Beliefs about the World The knowledge of the agent is represented as statements in \( W \) and includes the axioms and definitions that are needed to express the user model as well as information about the context. We may also require certain safety requirements to be included in the agent’s knowledge base. All the information in \( W \) will be included in every extension of \( T \), so all advice the agent can give will be consistent with this. We can include beliefs \( b \) about the world by introducing a default rule \( \langle c \cdot b \rangle \), where \( c \) describes the context that this belief is valid in.
Possible Actions  The possible actions of the user need to be expressible as a statement in the language of the knowledge base $W$. The actions which are contained in the extension $E$ will constitute the advice of the agent.

Behavior Patterns  Each behavior pattern is a combination of a context and the action that is taken in this context. It will generally not be possible for the user to follow all of these behaviors simultaneously. Instead, we regard the set of patterns to be a contextualized collection of the user’s possible actions. Each behavior pattern is formalized as a

<table>
<thead>
<tr>
<th>Concept</th>
<th>Example</th>
<th>Formalization</th>
<th>In</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge of the World</td>
<td>Only one action can be scheduled at the same time</td>
<td>Plan($t$, $a_1$) $\rightarrow \neg$ Plan($t$, $a_2$)</td>
<td>$W$</td>
</tr>
<tr>
<td></td>
<td>Each action can only be scheduled once a day</td>
<td>Plan($t_1$, $a$) $\rightarrow \neg$ Plan($t_2$, $a$)</td>
<td>$W$</td>
</tr>
<tr>
<td>Beliefs about the World</td>
<td>Unless stated otherwise, we assume the friend does not have time</td>
<td>$\neg$ Friend($r$)</td>
<td>$D$</td>
</tr>
<tr>
<td>Possible Actions</td>
<td>The user can use the morning to jog, go to the gym or read and use the afternoon to jog, read or get coffee.</td>
<td>Plan(Morn, Jog), Plan(Morn, Gym), Plan(Morn, Read), Plan(Aftn, Jog), Plan(Aftn, Read), Plan(Aftn, Coffee)</td>
<td></td>
</tr>
<tr>
<td>Behavior patterns</td>
<td>Jogging and reading are possible in any context, going to the gym is only considered in the morning</td>
<td>$\delta_1$ : Plan(Morn, Jog), $\delta_2$ : Plan(Morn, Gym), $\delta_3$ : Plan(Morn, Read)</td>
<td>$D$</td>
</tr>
<tr>
<td></td>
<td>Going for a coffee with a friend is only possible if the friend is available</td>
<td>$\delta_4$ : Plan(Aftn, Coffee) $\delta_5$ : Plan(Aftn, Jog) $\delta_6$ : Plan(Aftn, Read)</td>
<td>$D$</td>
</tr>
<tr>
<td>Preferences</td>
<td>In the morning jogging is preferred over the gym</td>
<td>$\delta_2 &lt; \delta_1$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>In the afternoon coffee with a friend is preferred over reading</td>
<td>$\delta_5 &lt; \delta_6$</td>
<td></td>
</tr>
<tr>
<td>Goals</td>
<td>Exercising once a day</td>
<td>ExerciseOnce $\rightarrow \neg$ ExerciseOnce $\neg$ (Plan(Morn, Jog) $\lor$ Plan(Morn, Gym) $\lor$ Plan(Aftn, Jog))</td>
<td>$W$</td>
</tr>
<tr>
<td></td>
<td>The goal requires at least one type of exercise</td>
<td>$E_1 = {\text{Plan(Morn, Jog)}, \text{Plan(Aftn, Read)}}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_2 = {\text{Plan(Morn, Gym)}, \text{Plan(Aftn, Jog)}}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_3 = {\text{Plan(Morn, Read)}, \text{Plan(Aftn, Jog)}}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Example of a User Model in Ordered Default Logic
default rule $\frac{c}{a}$ which contains a description of the context $C$ as the prerequisite and the action $a$ as the consequence and justification.

Preferences In our user model we consider the user’s preferences on behavior patterns. These preferences are represented using the ordering $<$ on the default rules. As noted in Section 2, the ordering $<$ must fulfill certain requirements to ensure that we can find a consistent extension of the initial theory. This will be an important challenge when populating the user model, but is out of the scope of this paper.

Goals We take goals to be concrete, desirable and collectively achievable outcomes that the user intends to work towards and we require the agents advice to lead to these goals being reached. We include each goal in the knowledge base $W$ of our theory $T = (W, D, <)$ as a sentence $g$. This means that every extension of this theory must contain, and be consistent with the assertion that the goal has been reached. As every goal is reachable, there are formulas $p_1, \ldots, p_n$ which represent the possible plans to achieve the goal. Each plan is a conjunction of actions which results in the goal $g$ being achieved. We include a statement $\neg(p_1 \lor \cdots \lor p_n) \rightarrow \neg g$ in $W$ which infers that the goal is not reached if none of the corresponding plans have been executed in an extension. If this occurs, the extension is inconsistent and will not be considered when providing advice to the user.

Advice The agent gives the user advice of which actions to perform. These suggestions are based on the action sentences which are contained in an extension $E$ of the theory $T$. If there are multiple consistent extensions, the agents needs a way to choose from these.

4. Human-Agent Realignment via Updates

Our motivation for using default logic to represent the agent’s user model was that this allows the user to interact with and adapt the agents reasoning process directly if the agent’s advice does not match the needs of the user. In the following, we refer to these situations as misalignment scenarios.

The three causes for these misalignments that are differentiated in [8] are the reasoning process of the agent being wrong, the agent’s user model being wrong, or something having changed in such a way that the agent needs to adapt to the new situation. The last case is further divided according to the concepts that could change, namely the context the user is in, the user’s internal state, and the user’s desired behavior. An overview of the different misalignment scenarios, including an example of what this scenario could be for the example agent we have introduced in Table 1, can be found in Table 2.

For the purpose of this work, we assume that the agent can accurately determine which misalignment scenario is causing the mismatch between between the agent’s support and the user’s expectations by communicating with the user. An example of how to design such an interaction between the agent and the user can be found in [8].

In order to resolve each of the identified misalignment scenarios, we now give explicit realignment updates. An overview of these updates is also included in Table 2.

**Example 3** In Table 3 we present which realignment updates correspond to the misalignment scenarios from Table 2 and the result of this update.
Table 2. Types of Misalignment Scenarios and Corresponding Realignment Updates

<table>
<thead>
<tr>
<th>Cause</th>
<th>#</th>
<th>Scenario</th>
<th>Example</th>
<th>Realignment Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incorrect World Model</td>
<td>1</td>
<td>Incorrect Knowledge</td>
<td>An action can be scheduled multiple times a day</td>
<td>$W \div \varphi$ or $W \ast \varphi$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Incorrect Beliefs</td>
<td>Going to the gym is not possible</td>
<td>$D \div \delta$ or $D + \delta$</td>
</tr>
<tr>
<td>Incorrect User Model</td>
<td>3</td>
<td>Incorrect Goals</td>
<td>The user plans to read every Saturday</td>
<td>$W \div {g, P}$ or $W \ast {g, P}$</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Incorrect Preferences</td>
<td>In the morning the user prefers going the gym rather than jogging</td>
<td>$&lt; \div \delta &lt; \delta' \ast \alpha$ or $W \ast \delta &lt; \delta'$</td>
</tr>
<tr>
<td>Change in Context</td>
<td>5</td>
<td>Incorrect Context</td>
<td>A friend is available for coffee</td>
<td>$W \ast c$</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>New Context</td>
<td>The user does not want to go jogging when it is raining</td>
<td>$W \ast {c, D} \div \frac{\varphi}{\varphi}$</td>
</tr>
<tr>
<td>Change in the User's Internal State</td>
<td>7</td>
<td>Certain Change</td>
<td>The user is too sick to exercise</td>
<td>$W \ast \varphi$</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>Add Possibility</td>
<td>The user wants to know whether jogging can be avoided</td>
<td>$T \div 1 \varphi$</td>
</tr>
<tr>
<td>Change in the User's Desired Behavior</td>
<td>9</td>
<td>Change of Goals</td>
<td>The user plans to read every Saturday</td>
<td>$W \div \varphi / W \ast \varphi$</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Incorrect Preferences</td>
<td>In the morning the user prefers going the gym rather than jogging</td>
<td>$W \div \varphi$ or $W \ast \varphi$</td>
</tr>
</tbody>
</table>

**Incorrect World Model** We understand misalignments due to the agent’s reasoning being wrong to manifest as mistakes in the knowledge and beliefs of the agent about the world. If the agent has incorrect knowledge of the world, this requires an update on the knowledge base $W$. The operators $\div$ and $\ast$ can be used on the set $W$ to either remove incorrect information or update new knowledge as introduced in 2.

If the beliefs in the agent’s world model are incorrect, this can be resolved by updating the set of default rules $D$. For a wrong belief, we first need to identify the default rule $\delta$ that this belief corresponds to and then remove it using the contraction operator $\div$ on $D$. If a belief is missing it can be added to the world model as a default rule $\delta$ using the expansion operator $+$. As mentioned in Section 2, these updates also entail additional updates that are necessary for the ordering $\prec$.

**Incorrect User Model** Misalignments of the user model can refer to the information the agent has about the user’s goals or preferences. When updating the goals, this also need to include changes to the corresponding plans $p_1, \ldots, p_n$. If a goal $g$ is removed, this results in the new knowledge base $W' = W \div \{g, P\}$, where $P = \neg(p_1 \vee \cdots \vee p_n) \rightarrow \neg g$. If a new goal $g'$ is added then we obtain the new knowledge base $W' = W \ast \{g, P\}$.

The user’s preferences between behavior patterns are expressed in the ordering $\prec$. While we have not directly introduced update operators on this ordering, using the translation given in Section 2, the ordering is contained in the knowledge base $W$. This means
that we can use the update operators defined for $W$ to update the user’s preferences. We use \( < \vdash \delta < \delta' \) to remove and \( < \ast \delta < \delta' \) to include a new preference.

**Context Changes**  Context changes refer to a change of the situation which the user is in. Misalignments of this type can occur in two different ways. Firstly, it is possible that the agent is wrong about the context the user is currently in but generally knows which support the user requires in this situation. In this case the context information needs to be updated in the knowledge base $W$ in order to fix the incorrect world model.

Secondly, the agent has no knowledge of the context $c$ yet, therefore the context is not recognized and there is no information about how this context should be handled. We resolve this by expressing the behavior patterns which contain the new context as default rules of the form \( \vdash \ast \delta \) which we include in $D$ using the expansion operator. We also need to ensure that this default rule is placed in the ordering $<$ correctly.

**Internal State Changes**  The user’s internal state includes any emotional, physical or mental factors which may lead to the user wanting different support from the agent. We distinguish between different levels of commitment that the user has towards these updates. If the user is completely certain that a specific part of the agent’s advice should be changed, then the agent should be able enforce this. This can be done by updating the knowledge base $W$ to include the sentence expressing the action, or negation of the action using the revision operator $\ast$. However, this update may lead to the removal of sentences which are used to express or regulate the goals of the user.

<table>
<thead>
<tr>
<th>#</th>
<th>Update</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$W \times { \neg (\text{Plan}(t, a_1) \rightarrow \neg \text{Plan}(t, a_2)) }$</td>
<td>$E_1^\ast = { \text{Plan(Morn, Jog)}, \text{Plan(Aftn, Jog)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_2^\ast = { \text{Plan(Morn, Jog)}, \text{Plan(Aftn, Read)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_3^\ast = { \text{Plan(Morn, Read)}, \text{Plan(Aftn, Jog)} }$</td>
</tr>
<tr>
<td>2</td>
<td>$D \vdash \text{Plan(Morn, Gym)}$</td>
<td>$E_1^\ast = { \text{Plan(Morn, Jog)}, \text{Plan(Aftn, Read)} }$</td>
</tr>
<tr>
<td></td>
<td>$\text{Plan(Morn, Gym)}$</td>
<td>$E_2^\ast = { \text{Plan(Morn, Read)}, \text{Plan(Aftn, Jog)} }$</td>
</tr>
<tr>
<td>3</td>
<td>$W \times { \text{ReadAtLeastOnce, } \neg (\text{Plan(Morn, Read)} \lor \text{Plan(Aftn, Read)}) \rightarrow \neg \text{ReadOnce} }$</td>
<td>$E_1^\ast = { \text{Plan(Morn, Jog)}, \text{Plan(Aftn, Read)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_2^\ast = { \text{Plan(Morn, Read)}, \text{Plan(Aftn, Jog)} }$</td>
</tr>
<tr>
<td>4</td>
<td>$&lt; \ast (\delta_1 &lt; \delta_2)$</td>
<td>$E_1^\ast = { \text{Plan(Morn, Gym)}, \text{Plan(Aftn, Read)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_2^\ast = { \text{Plan(Morn, Gym)}, \text{Plan(Aftn, Jog)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_3^\ast = { \text{Plan(Morn, Read)}, \text{Plan(Aftn, Jog)} }$</td>
</tr>
<tr>
<td>5</td>
<td>$W \times { \text{Friend(Aftn)} }$</td>
<td>$E_1^\ast = { \text{Plan(Morn, Jog)}, \text{Plan(Aftn, Coffee)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_2^\ast = { \text{Plan(Morn, Read)}, \text{Plan(Aftn, Jog)} }$</td>
</tr>
<tr>
<td>6</td>
<td>$W \times { \text{Rain} }$</td>
<td>$E_1^\ast = { \text{Plan(Morn, Gym)}, \text{Plan(Aftn, Read)} }$</td>
</tr>
<tr>
<td></td>
<td>$D \vdash \text{Rain} \vdash \neg \text{Plan(L,Jog)} \delta_1 \delta &lt; \ast \delta \ast \delta &lt; \delta$</td>
<td>$E_2^\ast = { \text{Plan(Morn, Gym)}, \text{Plan(Aftn, Jag)} }$</td>
</tr>
<tr>
<td>7</td>
<td>$W \times { \neg \text{Plan(Morn, Jog)} \land \neg \text{Plan(Morn, Gym)} \land \neg \text{Plan(Aftn, Jog)} }$</td>
<td>$E_1^\ast = { \text{Plan(Morn, Read)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_2^\ast = { \text{Plan(Aftn, Read)} }$</td>
</tr>
<tr>
<td>8</td>
<td>$T \vdash_1 { \neg \text{Plan(Morn, Jog)} \land \neg \text{Plan(Aftn, Jag)} }$</td>
<td>$E_1^\ast = { \text{Plan(Morn, Jog)}, \text{Plan(Aftn, Read)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_2^\ast = { \text{Plan(Morn, Read)}, \text{Plan(Aftn, Jag)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_3^\ast = { \text{Plan(Morn, Gym)}, \text{Plan(Aftn, Jag)} }$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$E_4^\ast = { \text{Plan(Morn, Gym)}, \text{Plan(Aftn, Read)} }$</td>
</tr>
</tbody>
</table>

Table 3. Realignment Updates corresponding to Examples # 1 - 8 in Table 2
If the user would prefer different support $\varphi$ but is still open to accepting the original suggestion, then this possibility should be introduced without being enforced. We can ensure that after the update there is an extension which aligns with the user’s changes, but this extension may not be the optimal extension according to the user’s goals and desires. This update is achieved by using the operator $+_1$ on the theory $T$, which preserves all previous extensions but adds at least one which contains the advice expressed in $\varphi$.

**Desired Behavior Changes**  
We do not treat this case separately but refer to the case of an incorrect user model.

### 5. Discussion

Our goal was to create a user model in a way which allows the user to interact with and adapt the agents reasoning process directly. We approached this by representing each part of the user model explicitly in a theory of ordered default logic and presented the different revisions that can be used for this.

The ordered default logic we have used can be translated into standard default logic, which means that existing theorem provers for default logic can be used to implement our framework. However, before this can be used in a behavior support agent there are still a number of issues to resolve. Most importantly, there needs to be a control mechanism to ensure that all changes that are made to the agent’s reasoning maintain the effectivity of the agent and the safety of the user.

We have already mentioned the need for a dialogue which can be used to determine the cause of the misalignment from the perspective of the user [8]. However, for these interactions to be effective, the agent should also be able to communicate the information it has available and what its advice is based on. We therefore need to find ways to explain the agent’s reasoning in ways that are understandable to the user, such as the work in [15]. This may for example include expanding the logic to keep track of the inference steps which were taken to arrive at each conclusion.

Further work can also be done in expanding the user model of the agent to include additional concepts such as temporal aspects, probabilistic aspects, values, or norms as seen for example in [16,17,18,19]. Finding ways to reduce the complexity of the framework and ensuring that it can be scaled for more complex situations will be necessary before implementing a realistic agent.

Lastly, while we have proposed this framework as an alternative to data-driven approaches for the purpose of adaptable and explainable reasoning, we do not view these methods as mutually exclusive. Data-driven approaches can be especially useful for recognizing behavior patterns and learning about the preferences of the user. Eventually we hope to combine the strengths of both approaches and find ways to include information which was obtained from data-driven approaches within our logical framework.

### Acknowledgements

This research was partly funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, grant number 024.004.022.
References


This page intentionally left blank
Doctoral Consortium Papers
This page intentionally left blank
Exploring Human-AI Collaboration and Explainability for Sustainable ML

EYA BEN CHAABEN a,1

Université Paris-Saclay, CNRS, Inria, LISN

Abstract. The collaboration between Human-Computer Interaction (HCI) and Machine Learning (ML) can effectively address sustainability challenges by developing intelligent systems that enhance user behavior and reduce environmental impact. In order to promote energy efficiency and avoid AI-waste throughout the entire life-cycle of ML applications, developers need the ability to make environmentally conscious decisions from the start. In facts, we aim to create tools that help ML practitioners understand how a certain decision affects the environment and guide them in identifying suitable, more environmentally friendly models for their projects. By participating to the doctoral consortium, we contribute to develop a discussion on building and developing a more sustainable future.

Keywords. Human-Computer-Interaction, Sustainability, Machine Learning

1. Introduction

Sustainability has emerged as an overall concern for society [4]. While research on the intersection of Machine Learning and sustainability began already more than two decades ago [8], the rapid increase in the complexity of the computational systems we use on a daily basis created a new awareness for this pressing issue. According to researchers, ML can be a double-edged sword that has the potential to address sustainability issues while also contributing to the problem in a growing way [11].

Several research fields are involved in this effort. However, research in HCI, in particular, is at an important moment, with the potential to worsen or ease these serious concerns.

Over the years, the HCI community considered various approaches to addressing sustainability, beginning with working on individual behavior [5], then applying persuasive technologies [3], and finally arriving at a call for 'Green Policy informatics' that enables sustainable HCI to leverage a more traditional HCI skillset in addressing sustainability issues [1]. At the same time, the majority of work on sustainable ML and artificial intelligence (AI) addresses how to improve data collection, power sources, and infrastructures, as well as how to quantify and lower the carbon footprint [11] associated with developing and fine-tuning an algorithm [6]. We explore this subject from within communities, aiming to integrate the two fields to foster more sustainable decision-making. Specifically, we employ Human-Computer Interaction (HCI) approaches to incorporate explainability for sustainable ML.
sustainable practices throughout the Machine Learning (ML) model selection and ML life cycles.

Our ultimate objective is to create tools that help ML practitioners understand how a certain decision affects the environment and guide them in identifying suitable, more environmentally friendly models for their projects through these steps:

1. Develop a better understanding of how users plan, search for, and select ML models.
2. Develop new interactive methods for users to express, define, and develop relevant project descriptions using a human-computer partnership approach.
3. Develop new explainability approaches to allow intelligent systems to communicate alternative ML models in a more context-dependent manner, while taking resource cost estimates into account.
4. Develop and apply evaluation methods to determine the efficacy of the interaction from a human perspective.

2. Related Work

We start with briefly introducing current work in HCI and sustainable ML.

The perspective of HCI: Human-Computer Interaction is the study of how people design, implement and use interactive computer systems and how computers affect individuals, organizations and societies [9]. With that it studies how people interact with and through computers and aims to reduce the complexity and/or increase the power of interactions with them, to empower users to focus on their tasks, needs and habits [2,10].

The perspective of Sustainable ML: Within the sustainable ML and artificial intelligence (AI) community there are two main branches of work: ML for sustainability and sustainability of ML [13]. The majority of work on ML for sustainable focuses on building ML models to predict the impact of technology, such as renewable energy [7] or electric cars [12], on the environment. On the other hand, sustainability of ML addresses how to enhance data collection, power sources, and infrastructures as well as how to quantify and lower the carbon footprint associated with developing and fine-tuning an algorithm [6]. For this project, our focus is on the sustainability of machine learning. To achieve this, we employ Human-Computer Interaction (HCI) methods to promote more sustainable ML practices.

3. Research Questions and Challenges

Our work uses ML-experts processes and workflows as a baseline to develop decision support system. While the ultimate goal is to aid sustainable decision making for the general ML developer community, i.e. experts and non-experts, our first steps aims to understand and reflect expert knowledge.

The first part of the project address the challenge to assist ML practitioners in expressing the desired objectives and limitations that an ML model should address in a specific project. We aim to provide a guide to explore these alternatives based on bench-
marks and simulations to reduce the SustainML’s overall impact, avoiding unnecessary network architecture searches and simulations.

Our work therefore addresses three main questions in this context:

1. Does providing energy consumption feedback for each contributing factor of the model selection process reduce the energy consumption of the final model training?
2. Does providing recommendations, e.g., model parameters or domain data, support users to align their model better to their use case?
3. Does providing energy consumption feedback for each contributing factor of the model selection process impact ML experts behavior when selecting ML models?
4. Does providing energy consumption feedback for each contributing factor of the model selection process improve the sustainability awareness of ML experts in the short-term and long-term?

4. Framing the Intersection of Sustainability with HCI and Machine Learning

As the thesis title suggests, we’ve identified three primary themes: ”HCI,” ”ML,” and ”sustainability.” As a result, it is critical to first identify the various points of intersection among these themes, frame them, and establish the context in which my thesis fits.

4.1. Methodology

To create a framework that outlines different research areas where sustainability intersects with HCI and ML, we gather data from questionnaires with eight HCI experts, interviews with thirteen ML experts, and existing literature. This enables us to identify current trends and areas of focus that already exist and propose new directions sparsely represented across domains.

4.2. Results

According to the study, researchers are generally aware of the environmental impact of technology. HCI researchers demonstrated a thorough understanding of sustainability throughout the technology life cycle, focusing on material waste and end-user hardware consumption. This is consistent with current trends in sustainable HCI, emphasizing individual researcher responsibility and prompting a reevaluation of research objectives and methodologies. Traditionally, sustainable ML focused on model and hardware impacts, but recent developments show that ML experts are addressing broader sustainability issues. Classifying academic work by research stance helps to identify gaps and standardize approaches across disciplines. While some research areas have received extensive attention, others remain underexplored.

5. Understanding Existing ML Workflows

In today’s fast changing technological environment, understanding existing machine learning (ML) activities is critical. These workflows, which consist of a number of struc-
tured phases ranging from problem understanding through model deployment, provide insights on best practices, optimization strategies, and the complicated process of building efficient ML models.

5.1. Interviews

In this study, we examine how ML experts find, choose, use, and understand models in their everyday work. Our goal is to better understand the ML tasks from the user description and their methods to select ML models. With this knowledge, we intend to identify relevant characteristics of a ML project in terms of constraints and context to differentiate between different models.

Study procedure:
Thirteen ML experts (11 male, 2 female, avg. 8 years of experience [SD=3.1]) were recruited via email or direct contact from local research facilities. They participated voluntarily, acknowledging their rights and data usage under GDPR through informed consent. Data collection involved mixed-approach thematic analysis, with top-down themes derived from literature and interview questions.

5.2. Current ML practice

Regarding the impact of ML on sustainability, our findings indicate varied considerations among participants, with some actively mitigating environmental impacts, others monitoring their model’s carbon footprint, and some not prioritizing sustainability due to perceived insignificance. Evaluation of both model and infrastructure impact highlighted challenges in measuring carbon emissions and emphasized the need for comprehensive assessment, including hardware and water consumption in data centers. The trend towards larger ML models emerged as a sustainability concern, raising questions about the necessity and environmental impact of increasingly complex models. Additionally, the study explored different perspectives on the intersection of ML and sustainability, encompassing environmental, social, and economic factors.

The study on model selection strategies revealed various approaches among participants. A clear starting point was crucial, though participants differed in whether they focused on understanding the problem, analyzing data, defining goals, or leveraging existing knowledge. Insufficient training and evaluation data posed challenges for half of the participants. The integration of literature, Large Language Models (LLMs), and AutoML was vital, blending personal and collective resources for decision-making. Participants balanced performance with user understandability, trust in AI tools, and data compatibility. Interpretability and explainability were critical for model viability, with most participants valuing transparency alongside performance.

6. Supporting Sustainable Decision Making in ML Practice

In this section, we go deeper into the insights extracted from our previous qualitative studies and discuss the implications they have for design.
6.1. Design Implications

Based on our thematic analysis of the interviews, we extracted design implications for supporting ML developers to make more sustainable decisions throughout the ML-learning life cycle. In Fig. 1 we have included design considerations aligning with the ML-life cycle.

![Figure 1. Overview of Design Implications based on our qualitative investigations](image)

**Problem understanding**

**DI 1:** System that help ML expert to make more sustainable decision, should allow users to define their project needs and constraints. This can be achieved by presenting relevant recent work related to the problem, showing similar case studies for reference, decomposition of the problem into simple tasks and facilitating the visualization of the problem through diagrams or mathematical expressions.

**Data understanding & preparation**

**DI 2:** System that help ML expert to make more sustainable decision, should allow users to explore and understand the available and potential input data. Our interviewed ML experts usually begin their model selection process by examining the data. Such system should offer essential features like data profiling, data transformation capabilities, error analysis tools, and functionality to identify patterns and trends within the data. We need to develop visualizations and exploratory data analysis tools that generate at the same time feedback of key sustainability indicators, such as energy consumption patterns, resource usage trends, or environmental impact assessments.

**Modeling**

**DI 3:** System that help ML expert to make more sustainable decision, should allow users to have an overview of the model performance for a specific domain and data format with explainable results. Systems should display different key performance metrics for each tested model to let the user compare and evaluate the effectiveness of different models. It
should also offer detailed explanations for model predictions and outcomes, highlighting the factors influencing model decisions.

**DI 4**: System that helps ML expert to make more sustainable decision, should allow users to find and define more sustainable computing environments for their models. Users have to consider constraints such as limited resources and the need to optimize model performance within these limitations. To address this, the system should provide information about the energy efficient hardware options, optimizations strategies, hyperparameters tuning techniques distributed training capabilities, and model compression methods.

**DI 5**: System that helps ML expert to make more sustainable decision, should allow users to analyze pre-existing work (collaboratively). Our interviewees expressed the practice to review results and experiences of others, which gives them valuable insights and additional arguments for choosing a model based on its historical performance, results, user satisfaction, and energy consumption. To address this, the system should offer community forums and discussion boards where the users can share their experiences, insights, and recommendations regarding model applications and their sustainability implications or establish model repository with user ratings and reviews.

**Evaluation**

**DI 6**: System that helps ML expert to make more sustainable decision, should allow users to get relevant information related to their needs and goals of models that could help him to reduce its energy consumption. Users are often unable to obtain detailed insights into a model’s energy consumption without investing large efforts and often do not have the time to conduct deep studies. System should therefore help developers to access this information quickly. This includes information on a variety of factors that influence energy consumption, such as training strategies, optimizer selection, hardware compatibility, model architecture, and model compression techniques.

**Deployment**

**DI 7**: System that helps ML expert to make more sustainable decision, should allow users to adjust their model descriptions iteratively to reflect on the trade-offs between performance and sustainability impact. Users are often flexible about some model specifications, which however can be different depending on the project, application domain or outside parameters such as human supervision. Demonstrating various trade-offs between performance metrics and sustainable impact such as providing energy consumption implications of small improvements in accuracy, e.g. a 0.01% gain, can help developers to make more informed decisions.

7. Discussion and Future work

First, we want to gain a better understanding of ML developers’ workflows, specifically planning, searching, and selecting ML models. To help developers make environmentally conscious decisions from the start, we conducted interviews and observational studies. These insights are used used to develop tools, strategies, or interventions to help ML developers effectively integrate sustainability considerations into their work processes. Furthermore, we may discover areas where additional research or resources are required.
to address knowledge or practice gaps related to long-term machine learning development. Then we work on new explainability approaches for intelligent systems that communicate alternative ML models contextually while taking resource costs into account. Additionally, we want to create interactive visualizations for developers and intelligent systems to explore existing and energy-efficient ML models, thereby encouraging the use of environmentally friendly algorithms.

We are currently in the implementation phase of the prototype. We focus on establishing the frameworks for both the frontend and backend components for our system. In order to determine if and how our prototype supports and changes developers’ decision-making processes in comparison to current processes, a structured observation with 12 ML developers will be conducted to evaluate the final prototype.

8. Acknowledgements

I express my deepest gratitude to my doctoral supervisors, Prof. Wendy Mackay and Dr. Janin Koch for their invaluable guidance throughout my research journey. Lastly, I acknowledge the generous financial support provided by European Union’s Horizon Europe research and innovation programme (HORIZON-CL4-2021-HUMAN-01) under grant agreement No 101070408, project SustainML (Application Aware, Life-Cycle Oriented Model-Hardware Co-Design Framework for Sustainable, Energy Efficient ML Systems).

References


Abstract. The integration of artificial intelligence (AI) into various domains offers promising opportunities for innovation and growth. However, alongside these benefits, there is a growing recognition of the potential societal impacts and ethical concerns associated with AI deployment. Addressing these considerations requires a shift towards more human-centred approaches to AI development, emphasising the importance of understanding humans and societal dynamics. My doctoral research explores the concept of Human-Centred Artificial Intelligence (HCAI) especially within the context of smart manufacturing, where AI technologies play a pivotal role in optimising production processes. By facilitating collaboration between humans and AI systems, often referred to as hybrid human-AI intelligence, organisations can enhance productivity while ensuring worker satisfaction and safety. This research proposal outlines a comprehensive investigation into HCAI design and development, aiming to identify novel techniques for integrating human and AI expertise effectively. Through a combination of theoretical analysis and empirical studies conducted in real-world manufacturing environments, this research seeks to contribute to the advancement of HCAI and facilitate the responsible deployment of AI systems in manufacturing and beyond.

Keywords. Human-Centred Artificial Intelligence, Artificial Intelligence, Human-Centred Design, Smart Manufacturing, Human-AI Collaboration, Hybrid Human-AI Intelligence

1. Context

In recent years, there has been a strong integration of artificial intelligence (AI) into various domains and systems [2]. On one hand, AI holds promise of new types of applications and business opportunities [11]. On the other hand, the nature of AI as well as the domains of integration can even lead to undesired societal impact like unfair treatment or discrimination of minority groups or privacy violations [3][11][14]. Addressing these AI-related considerations requires innovative approaches to AI development with more focus on the human factors, and human-AI interaction in diverse application areas [27][30]. Human-Centred Artificial Intelligence (HCAI) is a term referring to various individual, societal, and ethical considerations related to the development of AI [2][5][37]. Following the long traditions of human-centred design (HCD), HCAI emphasises the importance of putting humans and various societal considerations in the centre of the development [2][16]. This includes the basic starting points of understanding user needs and the contextual factors of system design, as well as introduces new ones that are particular to AI as technology [2][8][16][20][24].
Manufacturing is one domain where the integration of AI is strong [6][13]. Demographical reasons and the increasing demand for improved production efficiency are steering the transformation within the manufacturing domain towards smart manufacturing [6][9][26]. In this trend, companies are facing significant technological advancements regarding digitalisation, data analytics, and automation, enabled by advances in AI and the availability of big data [7][14][27][32]. Manufacturing industry is under pressure for more effective production and associated challenge of redesigning both their business processes and their work organisation. Human-AI collaboration is one of the facilitators of more effective productivity in smart manufacturing. In this new form of interaction, humans are paired with AI to work towards common goal [39]. In this collaboration both entities bring their unique strengths and intelligence to enhance the productivity of the collaboration [6][9][10][32]. Therefore, it can be referred as hybrid human-AI intelligence as well [38]. My research aims to contribute to the advancement HCAI and hybrid human-AI intelligence. Specifically, I am investigating what HCAI design and development requires, and novel techniques for integrating human and AI expertise in a way that truly supports effectiveness and enhances worker satisfaction and safety. By combining theoretical analysis with empirical studies in real-world manufacturing environments, I aspire to develop innovative solutions that empower human workers and AI systems to collaborate effectively towards common goals. This research proposal outlines the theoretical framework, research methodology, and potential contributions of the study, followed by a detailed research plan.

The topic of HCAI is relatively novel, however, a large body of literature on topic exists. My work will be based on academic papers related to the main factors of HCAI, like HCAI design [2][3][5][17][20][22][25][27], explainability [1], transparency [18], responsibility [15], fairness [14], ethics [19], and human control [36][37].

In addition to academic papers, the industry [21][24], organisations [11], and designers [8][24] have participated to HCAI design and development by publishing toolkits, guidance, and information on the topic. These tools and recommendations are more practical, aimed for the AI designers and developers. These are going to be acknowledged in my thesis as well.

2. Objectives and research questions

The main objectives of my thesis research are to understand what should be acknowledged in the development of AI that is human-centred. I concentrate on the AI-related factors, investigating what impact they have on the user and the society. Based on this understanding, I propose tools and guidance for the companies in AI design and development, to support them to develop AI solutions that are human centred. In addition, I am set to investigate ways how to support and guide companies in ways that are suitable for different development practices. I am approaching the objectives of this doctoral thesis with two key research questions:

1. What are the considerations, challenges, and relevant building blocks regarding HCAI development?
2. How to facilitate the design of HCAI and human-AI collaborations?
By addressing these research questions, this doctoral study aims to contribute to both academic scholarship and practical integrations of AI for designers and developers. The findings of this research are expected to provide valuable insights for AI designers and developers to facilitate the development of AI solutions that are truly designed for the humans. This can promote the adoption of AI technologies and enhance the effectiveness and trustworthiness of AI systems in the face of AI-enabled future. I will build the answers to the research questions via five studies that support each other and form coherent understanding answering to the main research question. The different studies forming my thesis are described in the next chapter.

3. Studies, methods, and evaluation

I am combining different qualitative research methods to seek answers to my research problem [4][35]. I am using human-centred approach, so I will involve the target users to my studies when possible, and I will try to have at least one iteration rounds for each study. In addition, I am evaluating the proposed tools and guidelines with the target users [16][28].

STUDY 1 – Human-Centred AI Design in Reality: A Study of Developer Companies’ Practices

Despite a strong interest in HCAI in academic research, there is little research-based understanding of how the new AI-related requirements and principles manifest in practise in AI development. Hence, my first study has its objectives in understand how HCAI is currently addressed in AI development practises in developer companies. This exploratory research is motivated by the goal of forming new knowledge on how the human-centred design practices are realised in the context of developing AI applications. To gain in-depth understanding of their development practices and to understand how HCAI principles manifest in the current practices of AI development, I conduct an interview study of practitioners from 12 AI developer companies in Finland. Participating companies represent a variety of domain areas and company settings. The AI applications these companies develop vary from automated sensor-based applications to recommendation systems.

STUDY 2 – Towards a Human-Centred Artificial Intelligence Maturity Model

My second study is investigating how to realise HCAI when designing systems that utilise novel computational tools and require consideration of increasingly broad set of requirements, spanning from fairness and transparency to accountability and ethics. I am set to understand what kind of tools could be useful for the companies, so that they could be adoptable and adaptable to various development practises. The objective of this study is to support the HCAI development practices in companies for the humans to have AI solutions that are efficient, trustworthy, and safe. Hence, I propose a maturity model for HCAI (HCAI-MM). In this study, I present the first phase of the model development, in which the central building blocks of HCAI are specified. I then propose model’s first phase, in which the central building blocks of HCAI are specified. Initial company requirements for the model’s structure and content are evaluated with semi-structured interviews with four AI developers.
Study 3 – Human-AI Collaboration in Smart Manufacturing: Concepts and Framework for Design

With the integration of AI and data in manufacturing, there is a shift from humans merely interacting with technology to actively collaborating with it, especially with AI-enabled agents. This shift brings changes in work practices, tasks, and organisations. I believe that it is important to understand the key concepts to acknowledge and address in the design of effective human-AI collaborations for smart manufacturing. As the objective of this study is to understand and identify what are the key concepts to acknowledge in the design of human-AI collaboration in smart manufacturing, I choose a scoping literature review as a study method. Based on the identified concepts, I propose a human-AI collaboration framework that offers an initial basis for the design of human-AI collaborative systems for smart manufacturing.

Study 4 – Human-AI Collaboration in Smart Manufacturing: Design Method and Guidelines

Human-AI collaboration in smart manufacturing is a growing trend, and this collaboration should be designed appropriately so that it truly effective and supports operator’s well-being. In study 4, I am set to understand what kind of design method and design guidelines could support manufacturing companies to design human-AI collaborations. Building on the identified key concepts and the framework of the study 3, I will develop a design method acknowledging the required HCAI factors and the key concepts to address in human-AI collaboration. I will approach the study objectives with interviews with company representatives and designers, in order to understand the user needs. I am evaluating the proposed methods and guidelines with real uses cases from manufacturing companies related to the research topic.

Study 5 – Human-Centred Artificial Intelligence Maturity Model for Smart Industry

HCAI and its requirements might be unfamiliar to manufacturing companies, as they are still finding their ways to work with AI. Main objectives of this study are to investigate how will adoption of AI transform work practises, the organisation, and the society, and to understand how organisations will need to change to accommodate a changing workforce with evolving preferences. By building on the preliminary maturity model from the study 2, I am aiming to build model that is appropriate to practical use and helps companies to reach a higher maturity in HCAI and to support HCAI practises in smart manufacturing. The model provides two perspectives:

1) "Organisational HCAI Awareness" – to develop HCAI maturity / HCAI awareness and understanding in the organisation
2) "HCAI in products and services" – product development capabilities to develop AI that is human-centred and HCAI ways of working in Smart Manufacturing

Objectives of this study is to support companies to increase their understanding and awareness in HCAI and to support then to reach higher by offering appropriate and concrete guidance how to develop their HCAI capabilities and that way their maturity regarding HCAI. The model will be based on the building blocks presented in the study 2. However, I am adapting that model to smart industry domain, and the domain related factors must be understood and included to the model. In addition, to build a maturity model that truly would be beneficial and useful for the companies, I must understand the company needs. Hence, I will conduct interviews with manufacturing companies and AI developers. The proposed tools I will evaluate with the AI developers and smart
manufacturing companies, to test its usefulness and effectiveness. The evaluation will be conducted via interviews to understand the designer and the company perspective for the proposed model.

4. Preliminary results

The timeline for my thesis is January 2022 – December 2025. Figure 1 shows the timeline of my thesis and steps taken and planned steps.

In study 1 I investigated companies’ views and practices regarding the development of AI applications, with a specific focus on human-centred viewpoints of the practices. Our study shows that the requirements of HCAI are not well acknowledged currently in AI development. In addition, AI development practices can lean heavily on the experience, methods, and tools familiar from regular SD practices even though data and uncertainty are AI-related factors that require new capabilities, roles, and tasks. These are the real-life struggles we are aiming to create guidance and support for the companies with the following studies. Hence, there is a need for a usable and informatic tool that can be used to support HCAI development in companies.

In study 2, I reviewed related literature and identified the following HCAI factors to acknowledge in HCAI development: (i) work with AI uncertainty, (ii) user control and human-AI collaboration, (iii) ethical development and use: transparency, accountability, fairness, and (iv) trustworthiness: explainability and transparency to build trust between the user and AI. These factors will serve as a building block of the model created in study 5. In addition, the results of the interviews showed that companies prefer short, compressed information in a check-list form with a possibility to get more information on the topic if needed. They also hope to get targeted tasks and tools related to certain AI product or projects. These insights serve as a starting point for the study 5.

In Study 3, I reviewed related literature and identified six key concepts related to the human-AI collaboration in smart manufacturing: goal, agent team, skills, task, system, and communication. Based on these concepts, I proposed a human-AI collaboration framework that presents the key concepts and their relations and offers an initial basis for the design of human-AI collaborative systems for smart manufacturing. These concepts serve as a base for developing the human-AI collaboration guidelines in study 4.

5. Discussion

The integration of AI into various domains has been accelerating in recent years, promising new applications and business opportunities while also raising concerns about potential societal impacts such as unfair treatment, discrimination, and privacy violations.
This integration extends deeply into the manufacturing industry, where the pressure for more effective production has catalysed a transformation toward smart manufacturing. Human-AI collaboration emerges as one of the key facilitators in smart manufacturing. This symbiotic relationship, often referred to as hybrid human-AI intelligence, holds the promise of revolutionising the manufacturing landscape by augmenting human capabilities and improving operational efficiency. As companies embrace technological advancements facilitated by AI, there emerges a critical need to ensure that these systems are developed with a human-centric approach, considering the well-being of end-users, and addressing ethical concerns. HCAI offers a framework for addressing these challenges by emphasising the importance of placing humans and societal considerations at the core of AI development. As we explore deeper into the development of HCAI and hybrid human-AI intelligence in the manufacturing sector, several key considerations appear. Firstly, there is a need to develop innovative techniques for integrating human and AI expertise seamlessly. This involves designing AI systems that complement human skills and adapt to diverse working environments, thereby fostering a harmonious collaboration between humans and machines. Additionally, ensuring the effectiveness and safety of human-AI interactions requires comprehensive training and support mechanisms for workers to navigate these novel work arrangements successfully. My research aims to address these challenges and contribute to the advancement of HCAI and hybrid human-AI intelligence in manufacturing. By combining theoretical insights with empirical studies conducted in real-world manufacturing environments, I seek to develop actionable solutions that empower human workers and AI systems alike. Through a holistic approach encompassing HCD, ethical considerations, and technological innovation, I aim to pave the way for a future where humans and AI collaborate synergistically, driving sustainable growth and innovation in the manufacturing industry. By prioritising workers well-being, ethical considerations, and effective human-AI collaboration, we can harness the transformative potential of AI while mitigating its associated risks, thereby creating a more inclusive, sustainable, and productive future for all stakeholders involved.

5.1. Future work

Building on the insights of the studies included to my doctoral research, in the future it would be beneficial to develop comprehensive toolkit for the HCAI design in manufacturing domain. This toolkit should include instructions, tools, and concrete guidelines to support appropriate HCAI and human-AI collaboration to fully benefit the new technological systems and to offer positive user and customer experience. Objective is to map and understand HCD requirements in manufacturing and to understand how to support companies to understand and gather their clients’ needs for a more tailored, engaging experience for their customers.

Acknowledgement

Supervisors: Kaisa Väänänen and Thomas Olsson, Tampere University
Funding: Study 1 and study 2 are funded from European Regional Development Fund, Cohesion Fund and European Social Fund (fund number 306672). Study 3, study 4, and study 5 are funded from Business Finland -funding (grant 2451/31/2022)
References


M. Hartikainen / Investigations to Facilitate Human-Centred AI Development in Manufacturing


Developing Meaningful Explanations for Machine Learning Models in the Telecom Domain

Henry Maathuis
University of Applied Sciences Utrecht
ORCiD ID: Henry Maathuis https://orcid.org/0009-0002-5542-0478.

Abstract. This study aims to develop and assess Explainable AI (XAI) tools tailored for internal telecom end-users. It focuses on delivering meaningful explanations informed by design principles, cognitive biases, and human decision-making theories. The research explores customizing XAI for telecom use-cases to support internal decision-making, while considering user preferences captured through elicitation studies. As part of this PhD study, a prescriptive framework will integrate cognitive biases, design principles, and human decision-making theory to effectively communicate AI explanations to end-users. User studies will be conducted to evaluate the effectiveness of the prototypes following from the framework.

Keywords. Explainable AI (XAI), Telecom, Meaningful Explanations, Cognitive Biases, User Interface Design, Decision-making support

1. Context

Over the years, powerful Machine Learning models, particularly non-linear Deep Learning (DL) models have emerged and have proven to be highly effective in various real-world applications [1]. One of the key reasons for their success is that they can capture complex relationships and patterns that simpler, linear models such as Logistic or Linear regression cannot [2,3,4].

Unfortunately, the increased complexity of DL models comes at the expense of explainability and interpretability [5] due to the many parameters and the non-linear nature of these models. This is particularly problematic in fields such as healthcare, or finance, where incorrect decisions, potentially due to the lack of transparency and interpretability of such models, can have significant impacts on individuals. Similarly, in the telecom domain, inaccurate decisions can lead to widespread service disruptions, imposing costs on service providers while simultaneously impacting customers. Hence, the ability to explain a model's decision is crucial in such domains [6].

Despite this limitation, the performance gains of DL models in many real-world applications cannot be denied, and they continue to outperform the simpler, more interpretable models in many cases. In problematic domains where performance is the primary concern, deploying such high-performance models should be approached with caution [7].
As such, a rapidly evolving field of research has emerged coined Explainable AI (XAI). One of the aims of XAI is to enable users of DL models to understand the reasoning behind the model’s predictions. This can potentially lead to more trust in the model [8,9,10]. There are two main types of explainability tools for machine learning models: model-agnostic and model-specific. Model-agnostic tools like SHAP [11] and LIME [5] attempt to interpret the output of arbitrary models by identifying the importance of each feature in generating the output. They are also the most popular explainability tools in practice given their model-agnostic nature. On the other hand, model-specific methods offer a deeper understanding of how a particular model operates and the rationale behind its predictions. These explanations potentially leverage the internal workings of the model, such as its architecture, parameters, and decision-making processes, to provide insights into its behavior. Model-specific explanations can offer more detailed and precise explanations but are typically limited to understanding only the specific model (or class of models) being analyzed.

Unfortunately, most model-specific and model-agnostic explainability research is motivated by technical considerations, catering specifically to AI engineers. To ensure however that explanations generated by these tools are understandable and meaningful to a broader range of stakeholders, there is a pressing need to transition towards a more comprehensive Human-Computer Interaction (HCI) approach [13,14]. This shift acknowledges the importance of incorporating human-centered design principles into explainability efforts [15,16], thereby enhancing the usability and effectiveness of XAI tools. Meaningfulness is closely tied to language and communication, suggesting that the conveyed information should hold significance or purpose for the recipient. In HCI, the concept of meaningfulness varies among individuals based on factors like education, expertise, and contextual elements such as mood and task. Overall, the notion of what constitutes a meaningful explanation is subjective and contingent upon various individual and contextual factors [17].

Tailoring explanations is critical for generating meaningful explanations in XAI [18,19]. Research in this domain has found that tailored explanations lead to a higher degree of satisfaction and trust in the system than those obtained with generic explanations [20,21] and that providing personalized explanations to users leads to better performance on decision-making tasks [22]. Another study involves developing explanations that are specific to the user's domain knowledge and expertise. For example, a system may provide a detailed explanation to a domain expert but a more simplified explanation to a novice user. An argument is made that explanations that are non-tailored can result in algorithmic aversion and various biases [23].

Prior research discusses the influence of cognitive biases on XAI-assisted decision-making [24]. While this study outlines principles to mitigate certain biases and links them to technical explanation types like feature importances or counterfactuals, it does not address the connection to tailored explanations. Additionally, in [25], user-centric explanations are generated, which connect explanations with the dual process model of human reasoning [26]. However, the authors did not provide guidelines for designing user interfaces to assist users in better understanding the inner workings of AI models.

As part of my PhD studies, I aim to design concrete guidelines for designing meaningful human-computer interactions. To this end, we develop a framework that integrates theories on cognitive biases, design principles and decision-making to offer specific guidelines on how to effectively visualize and communicate explanations of AI
systems to end-users. The framework will be evaluated by designing interaction prototypes for several telecommunication use cases. User studies will be conducted to evaluate the effectiveness of the prototypes. Requirements elicitation interviews will be conducted to further steer the development of the prototypes and to tailor the explanations to the wishes and needs of the users of the system.

2. Research Questions

My research goal is to develop and evaluate XAI tools that are meaningful for internal users of AI systems within the telecom industry. A meaningful explanation considers the wishes and requirements of the users through requirements elicitation. Additionally, a systematic literature review will be conducted to unify design principles, cognitive biases and human decision-making yielding a framework for designing explanations for users. The framework is used as input for the development of the XAI solutions.

The main research question are as follows: “How can Explainable AI be tailored for telecom use-cases to support internal decision-making while meeting user needs, and how can insights from design principles, cognitive biases, and human decision-making inform the generation of such explanations for stakeholders?”

The following six research sub-questions are addressed in different research phases.
1. Which cognitive biases hamper AI-supported human decision-making?
2. Which interface design opportunities enhance understandability of XAI explanations in the context of the selected cognitive biases?
3. How can theory from cognitive biases and user interface design be combined into a prescriptive framework for developing human-AI interactions for domain-expert decision-making?
4. How can human-AI interactions be optimized through requirements elicitation to ensure alignment with user needs?
5. What insights can be gained by applying XAI tools to specific use-cases within the telecom industry?
6. How can the insights gained from applying the XAI tools to specific use-cases in the telecom industry be used to further optimize and improve XAI tools?

To the best of our current knowledge, there is no prescriptive framework available that integrates theories on cognitive biases, design principles and human-decision making to offer specific guidelines on how to effectively visualize and communicate explanations of AI systems to end-users.

3. Research Challenges

A significant portion of the PhD research is dedicated to developing a prescriptive framework for Explainable AI (XAI). One of the primary challenges lies in conducting a comprehensive systematic review to identify common patterns among various papers detailing XAI system designs. This entails navigating through a vast array of literature to distill key insights that can inform the development of the framework. Once the
framework is established, another challenge emerges in its application for developing multiple XAI solutions tailored to the telecom industry. Furthermore, validating both the framework and the XAI solutions presents another hurdle, requiring a series of empirical studies to assess their impact on improving the understandability and usability of XAI explanations in real-world telecom scenarios. Extending the validation process to include potential use-cases from other sectors, such as finance or healthcare, adds complexity but also offers opportunities to demonstrate the framework's versatility and applicability across diverse domains. Overcoming these challenges will be essential to advancing the field of XAI and facilitating its practical implementation in various industries.

4. Method/Approach and Evaluation

Several research papers will be produced to address the research questions at hand. The first research paper which yields a prescriptive framework for developing meaningful human-AI interactions for domain-experts aims to answer the first three research sub-questions. To validate the framework, we assess its efficacy in real-world telecommunication use-cases. This assessment involves comparing prototypes developed from the framework with various baseline prototypes. Baseline prototypes may encompass screens displaying solely the model's outcome alongside a confidence score, as well as the presentation of raw, non-tailored technical explanations. The efficacy is measures specifically using both qualitative and quantitative measures. Amongst others, these could include performance on task, satisfaction of the explanation and trust in the explanation or system as a whole.

The first use-case focusses on a machine learning model designed to identify cable breakages and uncover their root causes by analyzing a vast array of alerts within a telecom network. This task places a significant emphasis on the operator's role, as they are tasked with interpreting these alerts to distinguish potential cable breakages from benign activities like power outages or false positive alerts. Ultimately, it falls upon the operator to determine whether a technician should be dispatched for maintenance. Historically, this process has burdened operators with the tedious task of sifting through numerous alarms to differentiate genuine faults from transient or non-essential alerts. While rule-based methods have offered some respite, their efficacy is limited by the intricate and variable nature of alarm patterns, necessitating continual updates to expert knowledge and manual intervention. To facilitate the operators, an XAI-system will be developed to assist users in their decision-making with the goal of minimizing falsely sending out mechanics based on incorrect model predictions.

Internal users of the AI-systems will be interviewed to gain a better understanding of their practice and what their requirements and wishes are from an automated decision support system. Requirements elicitation studies will be conducted to answer the 4th research sub-question.

Together with the framework and the elicitation studies, prototypes will be co-created with the various stakeholders of the use-cases. The insights gained will be published to answer the 5th research sub-question.
The insights gained will be used as input for another case study and allow for fine-tuning the existing framework, answering the 6th research sub-question.

5. Discussion and Future Work

Table 1 outlines the project's overall timeline. The initial use-case has been selected, and the setup of the literature study is clearly outlined.

In 2024, the objective is to publish several papers, including a structured literature review presenting the framework. Additionally, two papers detailing technical XAI solutions for a specific use-case will be submitted. A requirements elicitation study will be conducted for the first use-case.

Moving into 2025, the plan is to conduct a user evaluation study for the first use-case and continue work on the second use-case.

By 2026, we anticipate discussing the results with stakeholders and evaluating the performance of the prototype for the second use-case. A technical XAI solution paper will also be submitted.

In 2027, the focus shifts to finalizing work on the second prototype and publishing a user evaluation study for the second use-case. Additionally, the PhD thesis will be completed and defended.

Table 1. Timeline of the project

<table>
<thead>
<tr>
<th>Year</th>
<th>Tasks</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>2023</td>
<td>Literature Study</td>
<td>Select Use-Case (DONE)</td>
</tr>
<tr>
<td></td>
<td>Identify Use-Case 1</td>
<td>Setup Literature Study Design (DONE)</td>
</tr>
<tr>
<td>2024</td>
<td>Requirements Elicitation Study Use-Case 1</td>
<td>Overview requirements and needs users Use-Case 1 (March 2024)</td>
</tr>
<tr>
<td></td>
<td>Low-BARRIER Conference Paper Literature Study</td>
<td>Submit Low-BARRIER Conference Paper (April 2024)</td>
</tr>
<tr>
<td></td>
<td>Develop Prototype Use-Case 1</td>
<td>Submit Literature Study (July 2024)</td>
</tr>
<tr>
<td></td>
<td>Evaluate Prototype Use-Case 1</td>
<td>Submit technical paper regarding Use-Case 1 (December 2024)</td>
</tr>
<tr>
<td></td>
<td>Discuss results with stakeholders and obtain insights</td>
<td></td>
</tr>
<tr>
<td>2025</td>
<td>Develop Prototype Use-Case 2</td>
<td>Submit user evaluation study paper regarding Use-Case 1 (TBD)</td>
</tr>
<tr>
<td>2026</td>
<td>Evaluate Prototype Use-Case 2</td>
<td>Submit technical paper regarding Use-Case 2 (TBD)</td>
</tr>
<tr>
<td></td>
<td>Discuss results with stakeholders and obtain insights</td>
<td></td>
</tr>
<tr>
<td>2027</td>
<td>Round up work on second prototype Work on PhD Thesis</td>
<td>Submit user evaluation study paper regarding Use-Case 2</td>
</tr>
</tbody>
</table>
Acknowledgements

My supervisory team comprises Dr. E. Postma (Tilburg University) as the doctoral advisor. Dr. D. Sent (Jheronimus Academy of Data Science: JADS) serves as my daily supervisor, and Dr. D. Kolkman (Utrecht University) is another member of my supervisory team. Additionally, for technical support in the development of machine learning models, I receive assistance from Dr. C. van Gemeren (University of Applied Sciences Utrecht), who is not formally part of the supervisory team.

My PhD position is research collaboration (ICAI Lab) between KPN which is a telecom provider in the Netherlands, JADS (Jheronimus Academy of Data Science) and Applied University of Utrecht, all of which play a funding role.

References

[6] Lipton ZC. The mythis of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. 2018 Jun 1;16(3):31-57. doi: 10.1145/3236386.3241340.


Towards Value Alignment for Opaque Agents Through Concept Analysis and Inter-Agent Value Modelling

Jack MCKINLAY
University of Bath

Abstract. Value alignment, ensuring that artificial intelligence acts in ways aligned with humans, is a key challenge in the deployment of AI systems and an important aspect in the development of hybrid intelligence. A notable issue in value alignment is the lack of precision around what value alignment actually is and how to evaluate its success. Another issue is the fact that multiple agents are involved in value alignment, but these agents' values may not be easily understood or expressed. In order for autonomous agents to readily adapt to these agents' needs, and to promote interpretability of the system, an evidence-based means to assess the value priorities of agents is required. In my PhD, I am proposing a precise definition of value alignment through a qualitative concept analysis, synthesising themes from across the literature to produce a cohesive description. In addition, I am developing a novel mechanism for inter-agent value inference to enable agents to assess other agents' values in an interpretable way. My PhD will generate a cohesive description of the value alignment problem, identifying sub-problems for further research, and build a means for autonomous agents to perform value inference in pursuit of achieving value alignment.

Keywords. Value Alignment, Value Inference, Agent Modelling

1. Introduction

Artificial Intelligence (AI) offers numerous benefits for the world, but hasty deployment without proper preparation presents significant risks regarding the safety and usability of the technology. Value alignment, broadly understood as the challenge of ensuring that autonomous artificial agents will act in ways aligned with humans when deployed in society, is crucial to mitigating these risks.

Values are essential concepts in shared decision-making for humans, being guiding factors shared between cultures [1], and differences in how we act on values are a common source of conflict. If we develop AI that does not also act towards these values in ways we find agreeable, it will lead to problems in their use. For example, if a loan-decision AI model prioritised profit as the owning bank valued, but failed to consider unfair discrimination or avoiding economic risk as other stakeholders like customers and regulators would value, the tool would be
potentially harmful or banned. Instead, our goal should be to use value alignment to create AI systems that cooperate synergistically, proactively, and purposefully with humans, amplifying instead of replacing human intelligence, and empowering human values.

Themes within the topic of value alignment range from issues of user trust in AI and reliable operation [2,3], to philosophical debates around the ethics of creating AI [4] and the treatment of AI systems [5,6,7]. Most literature within value alignment research either focuses on the question of which values to align AI systems with, or how to align AI systems with values; these are referred to as the normative and technical challenges, respectively [8]. My PhD focuses primarily on the technical aspects of the problem, specifically the problems of defining value alignment in a technical context, and inferring values held by agents in a system.

A recurring issue in the value alignment literature is the lack of a consistent definition on what value alignment means. Many authors borrow from Russell [9], describing the problem as “aligning AI with human goals/values/preferences.”, but this definition is flawed. The verb ‘align’ does not convey a precise criterion, and the inconsistency regarding whether the authors intend to align with goals, values or preferences is problematic as these terms are not interchangeable [8,10]. Other themes that repeatedly occur in definitions include AI acting ‘beneficially for humans’ [11,12] or ‘ethically’ [13,14,15], as well as mentions of behaviour guarantees [16,17]. What qualifies as beneficial or ethical is itself determined by individual values and ethical frameworks, while behaviour guarantees are difficult to prove. The lack of a consistent definition of value alignment is unhelpful, as it makes it more difficult to synthesise research and exacerbates disagreement among key decision-makers [18, p. 154].

Another issue in current agent models is the lack of a means for agents to reason about the values held by other agents in the system. Values serve an important role in coordinating behaviour in social contexts [19], and the absence of a mechanism for AI to infer these values and reason about them, a process referred to as value inference [20], could explain their perceived amorality. If we can endow AI with the capacity to model values and their relative importance in other entities, and how its own actions will impact these values, it will contribute towards improving its ability to act in morally positive ways by not ignoring others’ values in its decision-making. In addition, being able to express these beliefs used in the reasoning process will improve the system’s interpretability by providing evidence behind decisions.

In my PhD, I am investigating both issues. First, I aim to create a clear definition of value alignment by synthesising literature from across the topic. My goal is to provide researchers with a coherent definition for developing solutions in the field. In the second part of my PhD, I will use this definition to guide the development of my mechanism for modelling the values present in agents in a system, for the purpose of assessing alignment and interpreting agents and their decision-making.

1.1. Background

While values themselves cannot be directly observed, we can do so indirectly through their influence on the agents’ actions. Aler Tubella et al. [21], supported
by van de Poel’s work [22], explored value monitoring through their glass-box framework, which maps moral values into explicitly verifiable norms for the system, and uses these to constrain inputs and outputs. Schwartz’s model of values [23], the predominant definition in the literature, also supports our ability to model values. Through a cross-cultural survey of over 20 countries, he identified a circumplex model containing 10 motivational domains for values. This provided a conceptual formalism of the relationship between different human values, supported by empirical evidence. Brunswik’s lens model [24] was developed to examine probabilistic functionalism, the perceptive theory that cues are approximate indicators of underlying objects [25]. This suggests an approach for integrating perception and uncertainty into value inference, with the latter being cited as a necessity in safe AI development by Russell [26].

1.2. Research Questions & Challenges

My first research question reflects my goal of generating a precise definition of value alignment:

“How do we define value alignment in such a way that is practical and consistent with previous theory?”

This can be further divided into two sub-questions:

1. What are the practical components of a value aligned system?
2. What are the consistent elements of value alignment considered in previous research?

Generating a definition consistent with previous theory comes from my intention to synthesise the topic’s diverse interdisciplinary discussion. To date, reviews on value alignment have generally avoided examining the topic as a whole, instead focusing on sub-topics within the domain, and hence have not produced a definition that draws upon all areas of the field.

I consider a practical definition to mean one that is:

1. Quantifiable, to allow for implementation in a machine usable form.
2. Commensurable, in that the degree of alignment can be compared meaningfully between different systems.
3. Interpretable, in that the degree of alignment can be meaningfully attributed to the system’s behaviour.
4. Tractable, in that alignment can be measured and assessed in a practical time frame.

My second research question concerns the modelling of values in a system containing autonomous agents:

“How can we infer the values possessed by an opaque agent with reference to a predefined group of values?”

This can also be split into sub-questions:

1. How do we establish what values are relevant to an opaque agent?
2. How do we construct and validate an inference mechanism?
The reason for focusing on opaque agents, referring to agents whose reasoning process cannot be interpreted, is two-fold: to create a model that is agnostic to the reasoning process used; and because of the prevalence of opacity in AI systems \[27\]. By a predefined group of values, I mean a collection of human values identified as relevant in the given context and given a relative importance ordering. I use this predefined group because modelling values without a reference point would most likely lack semantic meaning in any values identified, which would limit the method’s interpretability. Identifying these values poses a normative challenge, and it is worth examining the processes involved in doing so to better understand value alignment overall, but this is not the focus of this project.

Existing work in modelling values in opaque agents usually does so through methods that encode humans’ preferences into the utility functions of autonomous agents \[28,29,12\]. These methods are problematic in that the humans modelled are usually opaque agents themselves, and utility functions are notoriously lacking in interpretability. Hence this creates a situation where it is difficult to determine what values have been encoded, and how effectively this has been done. Utility functions also imply an inherent commensurability between values, which often causes problems in practice.

My research instead intends to exploit the relationship between values and indicators within the problem environment to create an interpretable model for values held by agents. I will do this by grounding these values as features in the agent’s operating environment, and then infer preference orderings between values by connecting an agent’s actions to their impacts on these features. By establishing preferences, or lack thereof, between an agent’s values using orderings, I reduce the concern regarding the commensurability of values, as orderings lack the need for scalar weights. I also support interpretability by being able to justify assumed value preferences based on observed behaviour.

2. Methodology

My PhD project consists of two parts. The first consists of a structured literature review to construct a definition of value alignment using qualitative coding. The second involves developing a mechanism for interpretable value inference, leading to the assessment of value alignment. The definition obtained in the first part will be used to inform the development of the mechanism in the second part.

2.1. Structured Literature Review

I am constructing a definition of value alignment incorporating previous theory through a qualitative concept analysis. I am the sole coder, negating the risk of coding disagreement. Saturation will be assessed by a running average of new codes retrieved from papers, with saturation reached when a given threshold is not reached for several papers. The exact criteria will be assessed as part of the review. Codes are collected inductively and synthesised into themes, from which a definition for value alignment will be derived in line with our criteria. The nature of this definition will take the most appropriate form, based on the analysis output.
So far, I have retrieved literature from the Scopus database, which was chosen due to featuring peer-reviewed papers from non-predatory journals, and limited myself to English language papers. Papers have been screened to focus on contributions to the implementation of value alignment. I have excluded discussions of the moral status of autonomous agents and governance around ethical AI. I have also conducted a bibliography search within relevant papers to find additional material not captured in the initial search.

My initial search has identified 167 relevant papers, for which I have performed an initial coding to identify important themes in the value alignment space. A high-level consolidation of this is given in Figure 1. Note that the survey is ongoing and these themes are not final.

From this initial work, I have built a framework to continue analysis and expand on my identified themes. This will involve further screening papers to select those that can best develop the codes of interest. I anticipate the review being completed by the summer.

2.2. Value Inference in Opaque Agents

I am developing a mathematical model to explicitly encode agents’ perspectives of values in a given context for the sake of modelling other agents’ values. My focus on agents’ perspectives is because the process of determining value alignment will require inferring the values held by other agents in the system, but an agent’s perspective of values will depend on their own ontology. I represent these perspectives as an agent’s orderings over predefined values and how they relate these values to features in the environment. Values are assumed to follow Schwartz’s definition as shared latent constructs that influence contextual behaviour and can be ordered [23]. The mathematical details are in development, and I intend to use graphical modelling to describe the relation between values and features. From here I will infer preference orderings between values based on observed actions. My choice of graphical modelling is inspired by the glass-box framework from Aler Tubella et al. [21]. It will promote interpretability in the system and also allow the integration of uncertainty through Bayesian modelling techniques. Incorporating uncertainty is done to improve the robustness of decision-making [30]. I illustrate my working model in Figure 2.

I treat values as latent variables that influence the agents’ behaviour to varying degrees, which leads to the agent impacting observable features in the system.
environment. How features are related to each value is based on the observing agent’s perception, which I call their lens. The lens gives contextual meaning to values by grounding them in contextual features and determines how the agent ascribes different value preferences to other agents based on what it observes. How these lenses might change between contexts is an important research question, but beyond the scope of the current work. For autonomous agents, I can control how features are related to values, but direct control is obviously not possible with humans. The output of this model is a probabilistic ordering of the observed agent’s values. By comparing these orderings with an agent’s own values, which can either be directly encoded or estimated through the agent’s lens, I will generate information that can be interpreted using our previously obtained definition of value alignment to make a judgement on the degree of alignment.

At this stage of my project, I am focusing on two-agent scenarios. This will allow me to focus on refining the inference mechanism without having to construct and manage an excessively complex simulation. Systems of practical interest will undoubtedly contain more than two agents, necessitating a deeper understanding of value aggregation mechanisms and promising future opportunities for this research. While I am not involving humans in this stage of the research, future studies will investigate the process of modelling humans’ value priorities by autonomous agents, and how this can be exploited to make value aligned decisions.

Acknowledgements

This research is supported by UK Research and Innovation grant number EP/S023437/1. Thank you to my supervisors Dr. Marina De Vos, Dr. Janina Hoffman, and Dr. Andreas Theodorou for their support in my PhD.
References


Neuro-Symbolic Approach for Tantrum Monitoring and Prevention in Individuals with Autism Spectrum Disorder: A Protocol for Virtual Agents

Andrea MONALDINI
Department of Computer Science University of Pisa, Pisa, Italy
Department of Information Engineering, Computer Science and Mathematics University of L’Aquila, Italy
ORCiD ID: Andrea Monaldini https://orcid.org/0009-0004-2518-2055

Abstract. Autism Spectrum Disorder (ASD) poses unique challenges, with individuals often experiencing difficulties in emotional regulation, leading to disruptive behaviors such as tantrums. This research protocol outlines the development of an innovative system designed to monitor and prevent tantrums in individuals with ASD, employing a neuro-symbolic approach. The proposed system integrates neural networks and symbolic reasoning to improve understanding and prediction of tantrum episodes. Leveraging real-time physiological and behavioral data, collected through wearable devices and user input, the system employs machine learning algorithms to detect patterns indicative of imminent tantrum events. Furthermore, a symbolic reasoning system interprets these patterns, taking into account individualized factors. The outcomes of this study are anticipated to contribute valuable insights to the growing field of technology-assisted interventions for neurodevelopmental disorders.

Keywords. Autism, Human-AI interaction and collaboration, User modeling and personalisation, Integration of learning and reasoning, Biosensing techniques

1. Context

1.1. Clinical problem

Autism remains an etiological mystery, its treatment has attracted the attention of many researchers who have used different techniques and technologies to improve the quality of life of ASD subjects.

In this regard, recent methodological advancements have focused on measuring emotions as organized patterns that emerge within interactions with others which may aid in better understanding the developmental mechanisms that underlie the nature, causes, and consequences of emotions—including tantrum. Machine learning is one of
the above-mentioned advances in computer science that can process data to highlight meaningful patterns. When artificial intelligence is used to investigate affective states of the human being, we enter the field of affective computing, defined as “computing that relates to, arises from, or deliberately influences emotion or other affective phenomena” and use “affect” and “emotion” interchangeably [16].

1.2. Related Literature

Most of the literature found with the keywords "ASD" and "Machine learning" refers to the diagnosis of the disorder itself, however the use of machine learning in tantrum detection in ASD children has been studied and applied in many researches (for a complete review: [10]), but only a few of these used biometric data: most used data related to facial expression, kinetics and motor movements ([15], [9], [8], [17]).

Studies that specifically deal with the detection of tantrums in ASD patients through the use of wearable devices and biometric data are now presented.

The first study [13] aims to examine the feasibility of detecting agitated behaviors in children with autism spectrum disorder (ASD) using data from wearable devices. Data were collected on a 9-year-old autistic child using the Empatica E4 device, which records electrodermal activity (EDA), pulsating blood volume (BVP), and acceleration (ACC). The study demonstrated the possibility of using data from wearable devices to detect the agitated behaviors of children with ASD during their daily activities. This could improve the quality of life of ASD patients and their caregivers by providing timely notification of such behaviors and preventing harmful consequences.

The second study [11] proposes and implements a non-invasive real-time deep learning-based Meltdown / Tantrum Detection System (MTDS) for ASD individuals. The system uses a prototype hardware in the form of a wearable bracelet, which records the heart rate, skin temperature, and galvanic skin response of individuals with ASD, and transmits the data to a server connected to the internet, where it is analyzed by deep learning models such as CNN, LSTM, and CNN-LSTM. The system can detect the state of anger or panic crisis in real-time and generate an email alert to caregivers or clinicians, and it also offers a graphical interface to monitor and analyze data and results. The size and sociodemographic and clinical characteristics of the sample are not specified. The results show that the CNN-LSTM-based system achieved higher accuracy (98) and lower absolute mean error (0.04) than the other algorithms.

The aim of the third study [19] is to create a device that can detect and prevent meltdowns in autistic children using stress sensors and machine learning algorithms. An optical sensor was used to measure heart rate and an electrodermal sensor to measure the galvanic response of the skin. An STM32-F446RE microcontroller converted analog signals to digital and transmitted them via Bluetooth to an Android application. The logistic regression model showed promising results with an accuracy score of 0.82 and a recall score of 0.83. The decision tree model showed a lower accuracy score of 0.76 and a recall score of 0.62. These results depend on the only child tested for this thesis and could change for other users depending on how their body reacts to stressful situations. The study has shown that meltdowns can be predicted using simple machine learning algorithms that don’t require large data to give generally good accuracy scores.

The aim of the latest study [7] was to evaluate whether physiological and movement data collected by a wearable biosensor can be used to predict aggression toward others in
young people with autism spectrum disorder (ASD). Cardiovascular, electrodermal and
accelerometric activity was recorded from a biosensor worn by 20 young people with
ASD (ages 6-17 years) during 69 nature observation sessions in a specialized psychiatric
unit. The results indicate that aggression towards others can be predicted 1 minute before
it occurs using 3 minutes of previous data from the biosensor with an average area under
the curve of 0.71 for a global model and 0.84. The biosensor was well tolerated, usable
data were obtained in all cases, and no users withdrew from the study. These findings lay
the groundwork for the future development of real-time adaptive intervention systems to
prevent or mitigate the emergence, occurrence, and impact of aggression in ASD.

As far as patient intervention is concerned, research so far has focused on robotic
agents in the field of social robotics: social robots, also known as Socially Assistive
Robots [5], have been designed to leverage their social and affective attributes to sustain
people’s engagement as well as to motivate, coach, educate, facilitate communication,
monitor performance, improve adherence to health regimen, and provide social support
to people [14]. Researchers have found that if a robot communicates using the same sorts
of non-verbal cues that people use, then people subconsciously interpret, form social
judgments, and respond to robots much as they do when these cues are used by people
[1].

2. Research Question

The research question is as follows: how does the integration of affective computing
with physiological signal processing, and a symbolic reasoning framework with Answer
Set Programming (ASP) contribute to the development of a virtual agent for tantrum
monitoring and prevention in individuals with ASD?

It is hypothesized that the virtual agent, incorporating advanced affective computing
techniques and a dynamic symbolic reasoning framework through ASP, will significantly
prevents or reduce the severity of tantrum episodes in individuals with ASD. Addition-
ally, it is anticipated that the personalized interventions generated by the virtual agent,
based on real-time emotional data and individualized factors, will enhance emotional
regulation and well-being, contributing to an improved quality of life for both individu-
als with ASD and their caregivers. Through iterative testing and refinement informed by
user feedback, it is anticipated continuous improvements in the virtual agent’s effective-
ness and user satisfaction over the course of the study. One of the possible explanations
for this clinical pattern is that individuals with ASD experience extreme difficulty filter-
ing sensory and social inputs, a fundamental process to discriminate important informa-
tion from irrelevant information when interacting with their environment (physical and
social) [20].

To talk about the novelty of this research, it is good to underline, starting from the
prevalence of robotic agents compared to virtual ones as therapeutic and educational
tools for autism, a particular phenomenon: the practices of the of the use of the robot
in the field of ASD, often considered best practices, are determined by the mainly from
manufacturers or professionals in the robotics sector who take little into the considera-
tion of clinical aspects [21]. The following research will be carried out with the collabor-
oration of a strongly multidisciplinary team that will ensure rigor both from a computer
science point of view and from that of social sciences and clinics. The second funda-
mental element of novelty is the use of a Neuro-Symbolic approach, which represents a convergence of symbolic reasoning and neural network based learning. This hybrid methodology leverages the strengths of both symbolic AI, which excels in rule-based reasoning and logical inference, and neural networks, which excel in learning complex patterns from data [6]. In the context of the presented research, the neuro-symbolic approach is particularly advantageous. It facilitates the integration of symbolic reasoning, crucial for modeling individualized factors influencing emotional states in ASD, with the adaptability and learning capabilities of neural networks. This synthesis enables the virtual agent to dynamically interpret and respond to complex emotional cues, evolving its interventions based on real-time data.

3. Methodology

3.1. Design

The design of this research is a cross-sectional study with quantitative and qualitative methodologies to investigate the development and impact of a virtual agent. This mixed-methods approach allows capturing both quantitative effectiveness metrics and the qualitative richness of user experiences. On the quantitative front, a pre-post intervention design will be implemented through a structured set of outcome measures, including frequency and severity of tantrums and identify the system’s effectiveness. Additionally, physiological data, collected through wearable devices, will be subject to statistical analysis to provide objective insights into emotional regulation changes. Complementing this quantitative foundation, a qualitative research strand will employ in-depth interviews and usability testing sessions. These qualitative components aim to uncover subjective experiences and perceptions of individuals with ASD and their caregivers regarding the virtual agent. Thematic analysis will be applied to extract rich narratives, providing a deeper understanding of how the virtual agent influences emotional well-being and daily life.

To develop a comprehensive understanding of the ASD population, a specific sampling strategy will be implemented. The sample will ideally be formed by subjects with a confirmed diagnosis of level 2 ASD between the ages of 18 and 30: by tailoring interventions to the specific needs and strengths of young adults, the effectiveness and long-term impact can be maximized. In tandem, caregivers closely associated with the daily lives of these individuals will also be invited to participate, providing a holistic perspective on the efficacy of the virtual agent proposed. To initiate participant recruitment, collaborative efforts will be made with the DISCAB Neuropsychology Clinic and the Abruzzo Regional Autism Reference Center. This collaborative approach, involving healthcare professionals and advocacy groups, will facilitate the identification and invitation of potential participants who meet the specific inclusion criteria. Informed consent will be obtained to confirm a clear and voluntary understanding of the purpose of the study. Finally, the determination of sample size will be guided by a balance between practical considerations and statistical power. Given the unique nature of the ASD population, an approach that prioritizes a smaller sample size with in-depth qualitative data collection will be adopted, ensuring a comprehensive exploration of the impact of the virtual agent on diverse individuals.
3.2. Neural network

In this section, we will explain how you intend to develop the ML system that will process the input data.

**Physiological Signal Processing Integration:** The biosignals that will be measured in the research have been chosen after the evidence highlighted previously in the review paragraph. Heart Rate Variability (HRV) is defined as the variations between consecutive RR intervals or heartbeats. HRV components are regarded as significant dependent measurements in psychophysiology and behavioral medicine, and they have attracted interest in both fields [2]. Galvanic Skin Response (GSR), also called Electrodermal Activity (EDA), is considered to be one of the most robust physiological indices of stress and it is widely used [12]. The aim is develop algorithms for HRV and GSR analysis to extract patterns associated with varying emotional states in individuals with ASD. It will be necessary to develop signal processing techniques to discern subtle changes indicative of heightened arousal or emotional distress.

**Real-time Emotional State Inference:** The system will integrate HRV and GSR signal processing components into a system for real-time emotional state inference. The aim is to provide the virtual agent with accurate and timely information regarding the user’s emotional state. Cognitive empathy is the capacity to understand and identify the feelings of another individual; it makes us better communicators because it helps us relay information in a way that best reaches the other person. It involves having the ability to comprehend the mental condition of others and what they might be thinking in response to the situation. This is associated with the concept known to psychologists as the “Theory of Mind,” or the capacity to consider the thoughts of others. Although they are far from the formalization of artificial empathy, several studies have highlighted how necessary it is for an agent to be able to intercept the emotional states of the user and respond in a relevant way [4].

3.3. Automated Reasoning

On the other side, this research will develop a comprehensive symbolic reasoning system that includes a broad understanding of individualized factors influencing emotional states in individuals with ASD. Automated reasoning is more complex than just analysing users’ behaviours since it requires problem solving based on numerous processes. The aim is to establish a knowledge base that incorporates symbolic representations of various contextual factors; this base will serve as the foundation for the virtual agent’s decision-making process, allowing it to interpret emotional cues within a rich and personalized context. To accomplish this, we plan to use ASP. This programming paradigm is widely used and has also been applied to the development of logical agents [3]: the reason is that, unlike neural networks, it allows for reliability, transparency, and trustworthiness. In addition, the absence of a black box makes it possible to intervene in a targeted manner to modify, if necessary, the system and ensure maximum customization. Using ASP’s expressive syntax will represent complex relationships and dependencies within the knowledge base, allowing for flexible and efficient reasoning. Another application of automatic reasoning within this research could be relative for mobility and transportation, as the cognitive profile of people with ASD can lead to some difficulties during their journey. This can involve difficulties with communication, the incapacity to
utilize public transportation securely, and frequently the requirement to be accompanied by a caretaker [18]. The problem of connecting two places as quickly as possible is a classic example used in ASP, which seems to have been created specifically to solve this type of problem. This element of the research will serve to provide ASD subjects with greater autonomy and reduce dependence on caretakers, providing overall a significant improvement in their quality of life.

4. Discussion and future works

The presented research poses several ethical considerations. Some potential ethical problems include:

1. Privacy and Informed Consent: Collecting sensitive data, including facial expressions, physiological signals, and behavioral patterns, raises concerns about privacy. Ensuring informed consent, especially considering the vulnerability of the ASD population, is crucial. In this regard, implement informed consent procedures in plane: clearly explaining the purpose of data collection, how the data will be used, and measures in place to protect participant privacy.

2. Stigmatization and Bias: A virtual agent designed specifically for ASD individuals may unintentionally reinforce stereotypes or stigmatize participants if not developed and implemented carefully. To avoid this, in this research they will be kept as a reference of inclusive and person-centered design principles, ensuring that the virtual agent’s responses and interventions are respectful and unbiased.

3. Equity and Accessibility: Ensuring equitable access to the virtual agent from diverse backgrounds is essential. Socioeconomic factors or limited access to technology could create disparities in the benefits of the intervention. For this reason, it is expected that the development of the agent and the affective computing system will be carried out through tools within everyone’s reach, such as common smartwatches or EMPATICA’s devices and monitoring platform.

By merging neural network and symbolic reasoning, this research aims to provide a novel and effective tool for supporting individuals with ASD in emotional regulation, fostering improved quality of life for both individuals and their caregivers. At present, the research is in the neural network development phase. Until now, the design of the agent has been based on data from the literature. The user’s interaction with the agent has not been specified in this paper because, for a design that is maximally user-oriented and context-specific, a survey is being completed to map the needs and expectations of young ASD subjects and their caregivers. These phases, carried out in parallel, are expected to be completed by July 2024; subsequently, with the data collected, we will move on to the development of the symbolic component: The idea behind this choice is the belief that similar interventions designed a priori have limited effectiveness. In conclusion, beyond the interdisciplinary synergy within IT science, this research underscores the imperative collaboration with psychology. It is in the confluence of IT science methodologies and psychology’s deep understanding of human behavior that we find the transformative potential to improve interventions in ASD.
References


Abstract. Medical imaging is a critical component of clinical decision-making, patient diagnosis, treatment planning, intervention, and therapy. However, due to the shortage of qualified radiologists, there is an increasing burden on healthcare practitioners, which underscores the need to develop reliable automated methods for interpreting medical images to reduce the time spent on commonplace cases and to support radiologists on more complex cases. Despite the development of novel computational techniques, automatically interpreting medical images remains challenging due to the subtlety and nuance of the patterns to be interpreted as well as the presence of noise and varying acquisition conditions. One promising solution to improve the reliability and accuracy of automated medical image analysis is interactive machine learning (IML), which integrates human expertise into the model training process. However, IML methods often lack compelling explanations to help users understand how a model is processing an image. To overcome this limitation, this study introduces a novel approach that leverages active learning (AL) to iteratively query for high-uncertainty samples while utilizing explanations from a prototypical part network to improve model classification. The proposed approach utilizes prototypical parts, which are snapshots of image sections, to determine an unlabelled image’s class based on the presence of the prototypical parts. Interaction occurs during the selection of prototypes and the AL phase, where a set of decision rules is designed to consider the contributions of which combinations of prototypical parts are the most representative of the unlabelled image output by the AL. The proposed explainable interactive machine learning (XIL) framework empowers medical experts to interact with the model’s training process, enabling more efficient and personalized learning through explanation and interaction.

Keywords. Human-in-the-loop (HITL), Interactive Machine Learning (IML), Active Learning (AL), Explainable Artificial Intelligence (XAI), Explainable interactive Machine Learning (XIL), Prototypical Part Network.

1. Context

The field of interactive machine learning (IML) has gained significant attention in the medical field in recent years [3,4,5,6,7]. Training image-based machine learning models typically relies solely on automated processes to learn patterns and make predictions,
offering no ability for interaction between the user and the model [7]. In contrast, IML incorporates human input and feedback into the modelling process, resulting in more interpretable and domain-specific models [3,4]. A typical IML workflow is presented in Figure 1, where the model training process is automated and periodically interspersed with interactions with a user. The user provides feedback to queries posed by the training process, which is then incorporated into another round of automated model training. Active learning [8], is a well known example of IML, and aims to reduce the number of labeled instances required to train machine learning models. However, there is an opportunity to deepen the interaction between the oracle1 by enabling the system to accompany user feedback queries with model explanations and allowing for more sophisticated feedback than simple labels in active learning.

The imperative for explainable artificial intelligence (XAI) in medical image analysis is driven by the need for transparency and understandability in model predictions [9]. Techniques such as local interpretable model-agnostic explanations (LIME) and shapley additive explanations (SHAP) have set benchmarks in demystifying model decisions by attributing predictions to specific features in the input data [10,11]. The integration of XAI into IML frameworks (XIL) ushers in an era where users can not only interact with models, but they can also grasp the 'why' behind model predictions. This is critical in medical diagnostics, where understanding the rationale for a particular diagnosis or treatment recommendation can significantly impact patient outcomes and clinician trust in AI tools.

Another emerging trend in XAI for medical imaging is the use of prototypes. Chen et al. [12] define a prototype as part of an image that is representative of features that capture the essential characteristics of a class. To prevent confusion with prior research that employs the term "prototype" [13], we adopt the phrase "prototypical part" to refer specifically to a portion of an image that captures essential characteristics, rather than full

---

1A human expert who actively participates in the training process of an Interactive Machine Learning (IML) model.
image prototypes. To arrive at a final classification for an unlabelled image, the evidence of multiple prototypical parts are combined.

Following this discourse, the AIMEE system [14] is one example of a family of XAI approaches that extract rules from models to help users understand how a model works, and even provide an opportunity to adapt a model. Systems like AIMEE work with numerical data only, however. There is though an opportunity to combine the use of prototypical parts with systems like AIMEE for models that work with medical images. To illustrate, we could have a rule such as "IF (prototype A) AND (prototype B) AND NOT (prototype C) THEN (Class B)". These rules could then serve to elucidate the current state of the learned model to a user, who could provide feedback to the learning process by adjusting the rules accordingly. Our work proposes to build such a system.

To complement the decision rules, we propose implementing interactive methods for training prototypical parts. The current approach outlined by Chen et al. [12] offers a relatively static methodology. However, not all prototypical parts may be readily interpretable by human specialists. For example, a model might generate a set of prototypical parts for specialists to either accept or reject. While rejected prototypical parts still hold value for the model, they may lack interpretability for human experts and consequently might not be incorporated into the Interactive Machine Learning (IML) framework. This setup would empower an oracle to exert influence in the training process by accepting or rejecting prototypical parts based on their representativeness of the class. This functionality offers an opportunity for the system to leverage the expertise of the human oracle, thereby ensuring the selection of more relevant prototypical parts.

The combination of these approaches AL, prototypical parts, and rule-based explanations—presents a holistic strategy for enhancing machine learning applications in medical image analysis. By prioritizing interpretability, efficiency, and expert integration, the framework being built in our work addresses key challenges in the field, offering a path toward models that are not only technically proficient but also clinically valuable.

2. Research Questions

This research will address the following research questions:

RQ1: When using AL for medical image classification problems, which combination of model (low, medium or high capacity), data representation (raw images or bottleneck features from a pre-trained model), and selection strategy (random, margin or Least-Confidence) leads to the most accurate models with the fewest labelled images?

When choosing an AL framework, several key factors come into play. Firstly, the nature of the dataset is crucial, as different AL strategies excel in different data contexts. For instance, uncertainty sampling may shine in image classification datasets, while ensemble-based AL could be more effective for text data. Additionally, the size of the dataset matters; some AL methods may require a larger initial labeled dataset to be effective, while others work well with smaller sets. Model capacity is also a consideration; high-capacity models may need more labeled data and time to converge, whereas simpler models might suffice with less data. Finally, the choice of sampling and query strategies—such as uncertainty sampling or query-by-committee—can significantly impact AL effectiveness.
In terms of classifier effectiveness within AL scenarios, it’s hypothesized that models capable of learning from few labeled examples are preferable. Pre-trained models like ResNet50 are anticipated to offer robustness but may be computationally intensive. To address this, an alternative AL framework is proposed, leveraging bottleneck features from ResNet50 combined with a Random Forest classifier. Additionally, shallow Convolutional Neural Networks (CNNs) are seen as promising due to their flexibility in architecture and training. This study aims to validate these hypotheses by analyzing various AL frameworks and classifiers, offering insights into optimal combinations for different dataset characteristics.

**RQ2:** How can interpretable decision rules based on prototypical parts (rather than feature values) be created to improve model interpretability?

In order to generate informative explanations, an XIL system will necessitate interpretable explanations, such as decision rules or rankings of feature importance that are understandable by humans. The process begins by translating the prototypical parts into a vector space. This enables a direct mapping of decision rules to specific prototypes or their combinations. The mapping would guide the network in emphasizing certain prototypes over others based on the decision rules. For example, if a decision rule indicates a particular feature is highly indicative of a class, the prototypes corresponding to that feature could be weighted more heavily in the classification process.

Additionally, the feedback mechanism from AIMEE, where users can edit or propose new rules, could be used to refine the set of prototypes uses. If a user identifies a prototype that does not contribute effectively to classification or misses a critical aspect, the network could be adjusted to incorporate this feedback, either by modifying the existing rule or learning new ones that better capture the user-defined rules.

**RQ3:** How can decision rules be modified by users based on the presence or absence of prototypical parts and incorporated as user feedback into an Explainable Interactive Machine Learning (XIL) framework?

In addressing this question, the focus is on enabling user engagement with decision rules in medical image classification within the XIL framework. Challenges include simplifying complex rules without losing effectiveness, designing intuitive interfaces, establishing effective feedback mechanisms, ensuring model interpretability, and techniques to incorporate revised rules back into a model. By prioritizing user interaction and overcoming these challenges, we aim to enhance the transparency and interpretability of machine learning models in medical image classification.

**RQ4:** To what extent can the integration of IML enable the discovery of more human interpretable prototypes?

In the initial pool of labeled data, prototypes are established to harness domain knowledge for the model. However, these prototypes, while valuable, may not always be interpretable to the user. To bridge this gap, a prototype ranking system is introduced, prioritizing user interpretability. Prototypes less clear to users remain important to the model’s functionality but are sidelined during user interactions. By enabling user interaction with the prototypes, allowing them to rank these based on interpretability, the model ensures that users are presented with prototypes that are meaningful and understandable. This strategy aims to balance domain significance with user interpretability, ensuring that engagement with decision rules always yields interpretable prototypical parts.
Table 1. The summary of results, measured by the Area Under Learning Curve (AULC), shows the best performing approach highlighted in bold

<table>
<thead>
<tr>
<th>Representation</th>
<th>Model</th>
<th>Sampling Strategy</th>
<th>PneumoniaMNIST</th>
<th>BloodMNIST</th>
<th>DermisMNIST</th>
<th>OrganMNIST</th>
<th>FractureMNIST</th>
<th>FractureMNIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottleneck Features</td>
<td>Random Forrest</td>
<td>Random</td>
<td>0.8270</td>
<td>0.6599</td>
<td>0.6469</td>
<td>0.6529</td>
<td>0.4151</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Margin</td>
<td>0.8295</td>
<td>0.7144</td>
<td>0.6786</td>
<td>0.6456</td>
<td>0.4109</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Least-Confidence</td>
<td>0.8284</td>
<td>0.6341</td>
<td>0.6789</td>
<td>0.6531</td>
<td>0.4273</td>
<td></td>
</tr>
<tr>
<td>Raw Image</td>
<td>ResNet50</td>
<td>Random</td>
<td>0.7960</td>
<td>0.9194</td>
<td>0.6892</td>
<td>0.9027</td>
<td>0.4329</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Margin</td>
<td>0.8614</td>
<td>0.9302</td>
<td>0.9119</td>
<td>0.9211</td>
<td>0.4283</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Least-Confidence</td>
<td>0.8492</td>
<td>0.9262</td>
<td>0.7134</td>
<td>0.9180</td>
<td>0.4542</td>
<td></td>
</tr>
<tr>
<td>Raw Image</td>
<td>Shallow CNN</td>
<td>Random</td>
<td>0.8213</td>
<td>0.7421</td>
<td>0.6562</td>
<td>0.7131</td>
<td>0.4088</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Margin</td>
<td>0.7879</td>
<td>0.6571</td>
<td>0.6597</td>
<td>0.7581</td>
<td>0.3922</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Least-Confidence</td>
<td>0.8335</td>
<td>0.7488</td>
<td>0.6602</td>
<td>0.7807</td>
<td>0.4073</td>
<td></td>
</tr>
</tbody>
</table>

3. Methodology

Early work has addressed RQ1 and designed an experiment aimed to assess the performance of pool-based active learning using various combinations of query strategies (random, margin, and Least-Confidence), model representations (raw image and bottleneck features), and model types (random forest, 5-layer CNN, and ResNet50). The raw image representations were resized to 224x224 and were used as input for the medium and high capacity models. To ensure a balanced representation across all classes, the AL workflow began by selecting an initial subset of labeled images consisting of 20 samples in all studies. Instead of relying on human agents for labeling, a synthetic approach was used to simulate adding a labels to images per iteration. During each of 240 iterations, the four most informative unlabeled instances were labeled and added to the dataset based on the chosen query strategy. The test set, which had already been split by the authors of MedMNIST [15], was used to evaluate the model’s generalization on unseen data at each iteration. Performance evaluation was based on two metrics: the area under the learning curve (AULC) and accuracy (ACC) after 100 iterations. The process was repeated for each combination of query strategy, model representation, and model type.

4. Results

This experiment aimed to determine the most effective combination of model representation, model capacity, and query strategies for active learning (AL) scenarios involving medical images. Our experiment involved two image representations (bottleneck feature and raw image representations) and three model architectures (Random Forest (low capacity), a 5-layer CNN (medium capacity), and a ResNet50 (high capacity)). We employed three query strategies, namely Random, Margin, and Least-Confidence, to identify the most informative data points for labeling.

The results, summarized in Tables 1 and 2, reveal that the high-capacity ResNet50 model using raw image representation, coupled with either the margin or least-confidence query strategies, consistently achieved superior performance compared to other combinations.

In summary, our study highlights the effectiveness of employing ResNet50 with raw image representations in AL scenarios. This approach achieves impressive accuracy while requiring significantly fewer labeled samples compared to benchmark models. Furthermore, our findings underscore the importance of selecting the appropriate query...
Table 2. The summary of results, measured by the Accuracy metric (ACC %), shows the best performing approach highlighted in bold.

<table>
<thead>
<tr>
<th>Representation</th>
<th>Model</th>
<th>Sampling Strategy</th>
<th>PneumoniaMNIST</th>
<th>BloodMNIST</th>
<th>DermMNIST</th>
<th>OrganMNIST3D</th>
<th>FractureMNIST3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottleneck Features</td>
<td>Random Forrest</td>
<td>Random</td>
<td>83.02</td>
<td>62.76</td>
<td>65.97</td>
<td>68.01</td>
<td>40.81</td>
</tr>
<tr>
<td></td>
<td>Margin</td>
<td></td>
<td>84.13</td>
<td>73.66</td>
<td>66.07</td>
<td>68.52</td>
<td>39.36</td>
</tr>
<tr>
<td></td>
<td>Least-Confidence</td>
<td></td>
<td>83.53</td>
<td>61.56</td>
<td>67.96</td>
<td>68.53</td>
<td>42.08</td>
</tr>
<tr>
<td>Raw Image</td>
<td>ResNet50</td>
<td>Random</td>
<td>79.81</td>
<td>93.74</td>
<td>60.02</td>
<td>87.51</td>
<td>43.79</td>
</tr>
<tr>
<td></td>
<td>Margin</td>
<td></td>
<td>87.82</td>
<td>96.66</td>
<td>72.15</td>
<td>93.12</td>
<td>40.41</td>
</tr>
<tr>
<td></td>
<td>Least-Confidence</td>
<td></td>
<td>86.70</td>
<td>96.34</td>
<td>72.76</td>
<td>92.39</td>
<td>45.09</td>
</tr>
<tr>
<td>Raw Image</td>
<td>Shallow CNN</td>
<td>Random</td>
<td>83.81</td>
<td>77.65</td>
<td>65.46</td>
<td>75.78</td>
<td>37.06</td>
</tr>
<tr>
<td></td>
<td>Margin</td>
<td></td>
<td>80.81</td>
<td>67.52</td>
<td>65.91</td>
<td>81.76</td>
<td>40.83</td>
</tr>
<tr>
<td></td>
<td>Least-Confidence</td>
<td></td>
<td>86.06</td>
<td>79.63</td>
<td>64.66</td>
<td>82.13</td>
<td>40.41</td>
</tr>
<tr>
<td>Benchmark</td>
<td>ResNet50</td>
<td>Random</td>
<td>85.70</td>
<td>95.60</td>
<td>75.1</td>
<td>85.70</td>
<td>40.40</td>
</tr>
<tr>
<td></td>
<td>Shallow CNN</td>
<td></td>
<td>83.20</td>
<td>79.42</td>
<td>66.54</td>
<td>81.91</td>
<td>40.12</td>
</tr>
</tbody>
</table>

strategy for optimal AL performance. The preferred strategy may vary depending on the image types and regions, as evidenced by our results. Moving forward, we plan to incorporate both Least-Confidence and Margin-based query strategies in future studies to ensure robustness and generalizability across different datasets and to investigate the use of different pre-trained model architectures, especially those targeted at medical images.

5. Future Work

The future work and completion plan outlined in this section focuses on addressing three key research questions related to improving model interpretability and incorporating human feedback in the context of medical image analysis. RQ2 aims to explore the creation of interpretable decision rules based on prototypical parts, rather than feature values, to enhance model interpretability. The plan involves integrating the ProtoPNET algorithm into the IML framework and adapting the AIMEE framework to generate decision rules based on prototypes. This approach aims to define clear conditions for prototypical parts in unlabeled images, thereby improving the interpretability of the model. For RQ3, the objective is to investigate how decision rules can be modified by users based on the presence or absence of prototypical parts and incorporated as user feedback into an XIL framework. The plan involves exploring how oracles can modify decision rules based on prototypes to build more trustworthy models using domain expertise. This approach aims to refine the model’s output to increase the expert’s understanding and establish trust during the decision-making process. Finally, RQ4 seeks to assess the extent to which the integration of IML can enable the discovery of more human-interpretable prototypes. This involves incorporating IML approaches into prototype discovery algorithms to involve human experts during the generation of prototypical parts. The aim is to identify more human-interpretable prototypes, thereby enhancing the applicability, trust, and interpretability of the models.

6. Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.
References

Designing Expressive Interaction with Generative Artificial Intelligence

Xiaohan PENG a,1
a LISN, Université Paris-Saclay, CNRS, Inria, France
ORCID ID: Xiaohan Peng https://orcid.org/0000-0002-5202-419X

Abstract. Current generative AI interfaces have posed challenges for creative professionals to interact with AI effectively. To address this issue from an interaction perspective, the primary focus of my research is designing expressive interaction with AI using mainly human-centered design approaches. More specifically, my research will delve into understanding how creative professionals use generative AI in professional settings, designing interactions with AI expressively and visually, and building engaging and controllable creative AI systems that keep a balanced agency between users and systems.

Keywords. Human-AI Interaction, Human-Centered Design, Creativity Support Tools, Expressivity

1. Context

Advancements in Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) have facilitated the production of high-quality synthetic images using descriptive text prompts [17,39], which has lowered the barriers to AI engagement for a wider range of audience including designers and artists. Current creative GenAI applications such as DALL·E 3 [35], stable diffusion [33], and midjourney [32] allow designers and artists to generate novel images, modify and iterate visual designs through text input. Research has shown that LLMs-based image generation can be especially useful in fast iteration and combining ideas in creative tasks [23]. Moreover, GenAI exhibits versatility across various creative fields by generating a range of outputs beyond images, such as music, videos, 3D models, and dance movements [47,37,1,9,29].

While current GenAI applications offer low-barrier text input access and great creative possibilities, non-AI professionals still often find it difficult to accurately express their intentions to AI and achieve their desired results [50,24]. More specifically, Subramonyam et al. have described the challenges with interacting with LLMs as intentionality gap (user only stating goals without other cognitive task processes), capability gap (user being unaware of AI’s capabilities) and language gap (user unable to express themselves clearly and effectively to AI) [44].
One of the most common technical solutions to these challenges is “prompt engineering”, which is a means of structuring text in order to efficiently interact with AI models and tune the generation process towards desired outcomes [48,41], for instance in text-to-image generation, using the structure like “A MEDIUM of SUBJECT in the STYLE style” and specifying explicitly color, technique, relationships in space, and motifs [26]. There are a large number of existing studies on prompt engineering that have been focusing on effective strategies for prompt writing [27,25,36,10,26], interactive prompt feedback [40,12] as well as more customized suggested prompt refinement [6]. Prompt engineering has also been studied in design practice. A study conducted by Chiou et al. on co-ideation between participants and AI image generators has demonstrated that strategies such as preparing data, breaking design into steps, crafting specific styles and reviewing can help generate effective prompts in design [10].

Apart from formulating effective prompts, some research uses conversational and turn-taking approaches in the interactions with GenAI, in order to share semantics between humans and AI and develop repairing mechanisms [44]. For instance, ChatPainter, a GAN-based model, uses a Q&A interaction such as “Q: is the woman standing on the board? A: no she is beside it.” to facilitate create image details that meet users’ expectations [43]. ControlNet allows users to use additional text descriptions to have incremental spatial control on existing generated images with diffusion models [51].

While these studies have provided insights into utilizing different AI input strategies to generate more satisfying results, the interaction remains text-heavy and always ends up in “trials and errors”. For visually oriented creative professionals, optimizing inputs in a rule-based way seems counter-intuitive, especially in divergent thinking processes [22]. Apart from prompt engineering and other techniques for detailing textual input, there seems to be a lack of studies on expressive and efficient approaches to empower designers and artists in interacting with GenAI.

In addition to inefficient and non-intuitive input interaction, I also see user engagement and control as another challenge in creative AI applications from a human-centered perspective. Despite GenAI has already been applied in a wide range of creative domains and practices such as quick exploration of visual ideas [46], UI design [34], fashion style clustering, forecasting and merging [18], speeding up design process [45], visual communication and fast prototyping [20,11], a large number of creative AI applications seem to be built in a fairly result-oriented way, where the results are presented but the creative process is missing. These tech-driven applications reflect an inclination within AI development to prioritize substitution over augmentation [46], and the absence of interaction makes them closer to “direct execution of goal” [44]. This potentially raises broader societal concerns such as deskilling [49]. On the other hand, the lack of engagement and control of the creative process also prompts users to view AI as an obedient tool. However, the relationship between human and AI can be rather versatile, and there can be different levels of agency in such interactions [21]. For instance, instead of functioning solely as a tool, AI can serve as a cooperative companion and assistant, guiding discussions and task outcomes [42]. Questions on how to maintain a balanced agency between humans and AI during interaction in context, and how to support such mixed-initiative and complex relationships remain to be answered.
2. Research Questions

In my research, the key research questions I am interested in are:

1. What are the interactions in users’ current creative practice, and how this can inform us to build a more controllable and meaningful system for professional design?
2. How to use expressive but controllable modalities and interaction techniques in designing interaction to support creative professionals to engage with GenAI?
3. How to integrate GenAI system interpretation and results meaningfully in design process to help users produce results more aligned with their expectations?

3. Methodology

In this section I will briefly introduce the theories and methodology I am likely to apply in my future research. I aim to use a mixed methodology in my research, where my primary methodology will be rooted in human-centered design, while also remaining receptive to theories and methodologies originating from other communities.

3.1. Generative Theories of Interaction as Methodological Building Block

In current human-AI collaboration research, a large number of methodologies applied seem to be tech-centered or highly reliant on existing technology. To ground my research from a human-centered perspective and envision interaction without being constrained by existing technology and existing usages of technology, one of the theoretical constructs that I will apply in my research is generative theories of interaction. A generative theory of interaction is grounded in empirically-based human activity and behavior theories, built based on concepts and generative principles related to the theories, and enables analytical, critical and constructive lenses to understand existing artifacts and inform future design [3]. In the context of HCI I see it can fit in my research both theoretically (as a theoretical foundation) and methodologically (using other theories as part of building blocks to inform design) since it constructs a generative bridge between HCI concepts and creating new designs. Two human-centered design theories that are closely relevant to my research and well-suited for the construct of generative theories are instrumental interaction [2] and human-computer partnership [30]. Since my research focuses on instant and immediate interaction between creative professionals and AI also new software adapting and appropriating patterns, these two theories will be great fits since they are suited for interactions in short-term and mid-term [3]. However, it might also be interesting to design for collaboration (although not being my main focus at the moment) and observe long-term community practices of using expressive AI interaction in the creative community, which might make community and common objects a relevant theory in a long-term project.

Considering using generative theory of interaction as a methodological building block, I would use “expressivity” as the main concept in my future research. Since in recent literature researchers have investigated the term “expressivity” in HCI context and found its meanings in sensories, dynamic forms, social activity and values [8], the level of concept abstraction should be appropriate, although the concept’s detailed definition

X. Peng / Designing Expressive Interaction with Generative Artificial Intelligence

411
might still require more in-depth elaboration. In terms of principles, initial thoughts are “input richness”, “expressive awareness” and “output variability”. “Input richness” and “output variability” are inspired by the principle of expressivity for human-computer partnership [30]. However, here I use two lower-level interpretations of the principle and they refer to the different levels of abstraction or resolution supported by input, and the capability of producing various and customized output based on different inputs. “Expressive awareness” is inspired by previous works on expressive communication devices [15,16]. From my perspective “expressive awareness” refers to the implicit sense of presence and connection during the interaction. The analytical, critical and constructive lenses that can be applied to the concept and principles remain to be researched from other existing literature.

3.2. Mixed Perspectives in Design

HCI methodologies that I apply in research will be primarily from a human-centered perspective. Depending on the stage of the research, different methodologies, mostly empirical will be applied such as semi-structured interviews, walkthroughs [28], thematic analysis and structured observation [14,5,19,22]. In the process of design, I will mainly design from third-person and second-person perspectives, meaning designing for users based on user studies and co-designing with users in context [13]. The decision on the perspectives is largely motivated by the discipline and the initiative to build more human-centered technology. Thus, design decisions are primarily informed by user perspectives or co-evolved in an engaging process. However, I am also open to using first-person methods such as research-through-design, which will be further deliberated depending on the subdomain of future research.

3.3. Learning and Getting Inspired from Users in Creative Practice

As an increasing number of GenAI applications built with different approaches are positioning at creativity support, one concern that comes into place is who the potential users will be for these new systems and how they will interact with them. From a human-centered perspective, I prefer to gather insights about users and draw design implications at the initial stage of design. In one completed project, a preliminary study has been conducted to understand how visually oriented users can interact with GenAI in their moodboard process, especially in terms of prompt formulation. The focuses are on the interaction and workflow with existing moodboard tools, how they prompt textually with GenAI, and their responses and actions towards the AI-generated output. The study included two think-aloud sessions, one of which was creating themed images with GenAI. An appropriated version of “Wizard of Oz” method [7] has been applied, which is instead of directly interacting with the GenAI, the experimenter played a mediator between the participant and the AI encouraging them to express their intents with technical impact minimized.

Apart from understanding users from a third-person perspective, more engaging activities such as participatory design workshops can be conducted to involve users in the design process [31]. In the scope of my research, participatory design is particularly suitable in the initial stage of exploring how artists and designers create professional work and how they are currently using GenAI in their practice. Activities such as co-
brainstorming, co-design with AI and finding creative usages of AI can be explored with various themes, such as the creative needs for precision or abstraction, their expectations of the roles of AI, their preferred ways of expressing intentions to AI.

3.4. Interaction in Human-AI Collaboration

To design interaction that is expressive but also controllable in professional creative settings, in my research design decisions are mainly drawn from relevant HCI theories, prototyping activities as well as empirical studies. Theories such as instrumental interaction [2], technical reasoning [38] and activity theory [4] used in HCI context work as an aid in creating a connection from thinking on objects of interests and actions to designing interaction. Prototyping as a process, particularly paper prototyping, prompts reasoning interactively, also enables a representation in physical form and helps form a rough model for interaction. Studies can also inform interaction design. In one of my completed projects, the results show that some designers consider “ease of steering” important thus they prefer using a system with simple interfaces and low cognitive workload; some designers think “system assumptions” affect controllability, thus they sometimes prefer manual control rather than let AI or the computer “make interpretations” for them; some designers think being able to control the input to AI precisely is the key to control, while some others interpret control as controlling over the whole AI generation process. These findings suggest that designers’ definitions of “control” over the AI affect their choice of interaction and inform design possibilities from each or combined implications.

4. Preliminary results

In the first project completed during my PhD research, I developed an interactive ideation system that allows designers to explore and express themselves when working together with GenAI tools using multimodal interaction. The system is built based on design implications obtained from the preliminary study, and it is digital web-based moodboard tool that incorporates GenAI functionalities with multimodal prompts including images, colors and semantics rather than pure text prompts. In comparative structured observation study with 12 professional designers, I compared the multimodal prompt moodboard tool with a baseline system that uses text prompts. The results have highlighted that prompt decomposition (from text to multimodal) allows users to explore underlying image connections and express intentions to GenAI. The work suggests design implications for designing expressive and controllable human-GenAI interaction in a design context and provides insights into the potential of using multimodal input to help designers better express themselves.

5. Discussion and Future work

Findings from my finished research project indicate that utilizing multimodal input rather than relying solely on textual input enhances designers’ ability to express ideas and explore diverse creative AI system usages. Given that my completed work incorporated images, colors, and semantics as AI input, future projects can delve into varied aspects which can include investigating the suitability of different modalities for creative AI in-
teraction in different use cases, assessing the stability and reliability of these modalities in AI interaction, exploring the expressive potential inherent in these modalities, and determining meaningful ways to combine them in future AI-incorporated designs. Apart from modalities, I am also interested in the potential of expressive and artistic mediums, e.g. drawing and sketching. In my future research, I would like to explore using drawing and sketching as an interaction technique or a part of design activity in collaborating with AI.

Implicated by different “control” interpretations, my future work will also dive deeper into specific controllability needs and navigate the balance between these different facets of control. In order to achieve this goal, part of my future work would focus on rethinking the interaction paradigm in human-AI from a user action point of view. I aim to use more expressive interaction techniques such as sketching to build a controllable system. Additionally, I am also interested in designing more fluid relationships and control between human and AI with different levels of agency, rather than using AI solely as a tech-driven obedient tool. In the long term, I envision my research potentially extending beyond GenAI interaction with creative professionals to include a broader demographic in their daily routines.

6. Acknowledgements

Wendy Mackay and Janin Koch supervise the author. The author’s PhD thesis is funded by ED STIC Université Paris-Saclay (D23-ET046) and the European Union’s Horizon 2020 research and innovation programme under grant agreement No 952026 (HumanE-AI-Net Project).

References


X. Peng / Designing Expressive Interaction with Generative Artificial Intelligence 415


[36] Hariharan Subramonyam, Christopher Lawrence Pondoc, Colleen Seifert, Maneesh Agrawala, and Roy


Interactive Multi-Objective Optimization for Military Helicopter Route Planning

Thomas QUADT

Tilburg University, Netherlands Defense Academy (NLDA)
ORCiD ID: Thomas Quadt https://orcid.org/0009-0009-4357-9781

Abstract. Information overload and the labor-intensive nature of the military decision-making process have made human decision-making deficient in modern warfare. To mitigate the current deficiencies, AI systems have been proposed. In a military context, these systems should be reliable and explainable, warranting a human-AI collaboration approach. One process that can greatly benefit from decision support is military helicopter route planning. It has many different conflicting criteria related to it, necessitating multi-objective optimization (MOO) techniques. In the literature, there is a lack of research on the decision-maker’s role. To fill this gap, interactive MOO is proposed to utilize the advantages of human-AI collaboration and create a shared understanding of the problem and its tradeoffs. Unlike the current interactive MOO research, the project is done with the active involvement of domain experts. Four different studies are proposed to investigate this problem. First, a literature study will be performed to review the state of the art. Next, an interactive MOO method will be designed, which will be applied to an Air Interdiction mission. Simultaneously, we will investigate the effect of the level of detail of threat modeling on the resulting optimization problem. Finally, the decision-support tool will be applied to a Strike Coordination and Reconnaissance mission.

Keywords. Route Planning, Mission Optimization, Interactive Multi-Objective Optimization, Human-AI Collaboration, human-in-the-loop

1. Motivation and Background

In 2020, the Dutch Ministry of Defence published its vision for 2035, stressing the need to become an information-driven organization [1]. One key planning process that can benefit from an information-driven approach is the Military Decision-Making Process (MDMP). It is used to facilitate creative, critical, and collaborative planning [2]. For the MDMP to be successful, decisions must be grounded in up-to-date, relevant, and timely information [3]. Yet, today’s increased unstructured data supply has complicated the operational environment of mission planning [2], resulting in information overload. Besides information overload, the MDMP suffers from its labor-intensive nature. It is heavily reliant on creating handcrafted products, making it a time-intensive production process [4]. This is problematic because, at the tactical level, military victory can be gained by outpacing the opponent’s OODA loop (Observation, Orientation, Decision, Action) [5], necessitating efficient planning processes [6]. Based on information overload, and

1 Thomas Quadt, t.j.quadt@tilburguniversity.edu
the labor-intensive nature, Meerveld et al. [7] argue that human decision-making mechanisms are deficient in modern warfare. In fact, military units have suffered a significant loss in their ability to conduct a detailed MDMP [2].

As a solution Artificial Intelligence (AI) systems have been proposed. Their expandable memory, processing speed, consistency, and logical reasoning power [8,7] make them highly valuable in supporting human decision-making and optimizing its outcome [9,10]. For example, they can greatly speed up the OODA loop, outpacing the opponent [11]. However, like humans, machines have their weaknesses. They are susceptible to a multitude of biases [12] as well as a lack of explainability [8]. This poses a significant challenge as NATO warrants a high level of explainability, responsibility, reliability, and bias mitigation for using AI within a military context [13]. Many of the drawbacks (e.g. lack of trust and explainability) of AI tools are characterized by the lack of collaboration with the system and decision maker (DM) [14]. Moreover, the complementary relationship between human decision-making and machine decision-making suggests the need for a higher level of human-AI collaboration [8]. Therefore, AI systems and humans working collaboratively with a shared understanding of the task, the context, and each other’s perspectives and capabilities can greatly benefit the MDMP [14].

A specific case in which the MDMP can benefit from AI-powered decision support is planning helicopter missions [10]. More particularly, helicopter route planning, which in its most simplistic form can be defined as finding the optimal route from A to B and back to A. The definition of an optimal route is shaped by different goals, mathematically represented as optimization objectives. Therefore, finding such a route is no trivial task. In the literature, nearly all methods include the objectives of minimizing route length and maximizing survivability (for surveys see: [15,16,17,18]). Even in trivial cases, these objectives become conflicting. For example, the shortest route could go straight through a hostile threat zone, whilst a safe route would take a longer detour and go around the threat zone. Besides safety and speed, studies can include the conflicting objectives of terrain masking [19], helicopter performance limits [20], mission effectiveness [21], pilot workload [22], and many more. All these conflicting objectives result in a possibly infinite set of optimal solutions instead of a single optimal solution. Therefore, to get a single optimal solution, the DM’s preferences need to be included in the decision-support.

The multi-objective optimization (MOO) literature [23] has identified four different classes of handling conflicting objectives based on the role of the DM in the optimization process. Firstly, in no-preference methods: no DM is present so a neutral compromise solution needs to be identified. In a-priori methods, a DM specifies preference information before the optimization process. A DM can, for example, specify the relative importance of safety and route length (e.g. (60%, 40%)). The advantage is that after linear aggregation, single-objective optimization techniques can be used to identify one unique solution. Note that even this single-objective optimization problem is often NP-hard [24]. The main limitation is that the DM’s limited knowledge of the problem beforehand results in inaccurate and misleading preferences [25,26]. Alternatively, a-posteriori methods can be used. These methods aim to approximate the entire Pareto Front and let the DM choose their preferred solution afterward. The Pareto Front is the set of solutions where no objective can be improved without degrading the performance of another objective. The resulting set (i.e. routes) can show, for example, that to improve the survivability with a single percentage a large increase in flight time is needed. The main limitation occurs due to the computational complexity of generating a Pareto optimal set of
solutions [27]. Moreover, when there are more than two objectives, the large amount of information required to be compared by the DM becomes problematic (i.e. information overload) [25]. When a DM is present to participate in the optimization process, these limitations can be overcome by using human-AI collaboration [28]. Interactive optimization methods enable a DM to guide the optimization process to its preferred region on the Pareto Front, drastically reducing the computational complexity [29,30]. At the same time, focusing computational power on preferred regions results in better approximations of those regions [31]. The iterative optimization process consists of identifying the preferred region, optimizing the Pareto Front in that region, and understanding the tradeoffs. It utilizes human-AI collaboration to find an optimal route preferred by the DM whilst simultaneously building trust and understanding of the problem, the tradeoffs involved, and the feasibility of their preferences [29,32,33].

To date, the research on military (helicopter) route planning has neglected the role of the DM. The literature has predominantly focused on a-priori optimization techniques [17] with pre-specified weights. Although less common, several studies have focused on a-posteriori optimization methods [34,35] but they rarely investigated how a DM should select solutions from the Pareto Front [31]. The only work on interactive optimization methods for military route planning was done in [27]. They discretize the world into a grid and create an interactive algorithm that generates a preferred region of the Pareto front, including the objectives of safety and flight time. So far, the algorithm has only been tested in the highly simplified setting of visiting five points avoiding three static known radars. Although they extend their work to a continuous terrain setting [36], the application remains simplified. Most importantly, they do not investigate the interplay with the DM but simply assume the user has a linear weighted preference beforehand. This trend follows a general research trend in the interactive MOO community: there is a strong focus on optimization, neglecting the DM [37,38]. For example, a survey assessing the performance evaluation of interactive MOO methods showed that only 4 out of 45 methods included domain experts in their experiments [29]. Therefore, it is no surprise that current methods do not sufficiently support the DM in finding the optimal solution and are often too cognitively demanding resulting in limited real-world usability [38]. This oversight is noteworthy because an increased focus on human-AI collaboration has the potential to mitigate the problems of human and machine decision-making thereby improving real-world usability of interactive MOO methods drastically.

2. Problem Statement and Research Questions

Considering a DM-centered human-AI collaboration approach, this Ph.D. project aims to examine the following research problem:

How can interactive multi-objective optimization help the Defence Helicopter Command (DHC) with military helicopter route planning?

To answer the problem above, performance metrics need to be defined. Traditionally, the research has mainly focused on assessing the performance of MOO methods based on algorithmic properties: the cardinality, convergence, and distribution of the Pareto Front [39]. Unfortunately, these metrics fail to capture the true goal of the decision-support tool. Namely, at its core, it is designed to aid the DM in finding their most preferred
solution. Consequently, properties concerning the DM: 1) the degree of confidence in the acquired solution, 2) the cognitive burden on the DM, 3) the extent to which preferences are captured, and 4) whether the DM has learned to understand the problem and its tradeoffs, need to be measured as well [29]. Moreover, Pour et al. [40] have shown that performance indicators specific to interactive optimization methods are needed since a-priori or a-posteriori performance indicators are not directly applicable [40]. Currently, only one such method exists [41] and it only measures algorithmic properties. In conclusion, the choice of metrics is a difficult one and an emerging research field. A definite list of performance indicators is thus not feasible at this stage of research but is in progress. Besides performance indicators, the military helicopter route planning problem and its modeling choices need to be defined, which introduces the first research question:

**RQ1:** What modeling choices are important for the military helicopter route planning problem and how does the state-of-the-art model them?

Based on the results of RQ1, an interactive MOO method will be developed (see Section 3). The resulting tool will first be designed to aid in an Air Interdiction mission. The Air Force defines Air Interdiction as air operations to disrupt, delay, divert, or destroy already located enemy military resources before they can be used effectively [42,43]. It happens so far from friendly forces that coordination with friendly movement is unnecessary. Since Air Interdiction is used on pre-located targets, it can be modeled as routing the helicopter from A to B and back. Within the military route planning literature this is the most investigated problem. Its many objectives are usually aggregated into safety and flight time, making it a bi-objective problem [17]. The novel application of interactive MOO methods for route planning, designed with the active involvement of domain experts, has given rise to the following research question:

**RQ2:** How does interactive MOO compare to a-priori, a-posterior, and manual route planning in the context of Air Interdiction missions?

Next, we will further investigate the level of detail of modeling required for the most important factor related to military route planning: threat modeling. In most cases, threats cannot be avoided entirely. Therefore, it is important to minimize the exposure to threats. The current literature has proposed many different methods of modeling threats. For example, the distance [44,45] or a function of the radar power [46,47] are used. At the cost of highly increased time complexity, more advanced methods like simulations can be used to estimate flight time and threat exposure as reliably as possible. Moreover, in realistic mission settings, the exact location and power of the threat are often unknown. Besides that, threats can be dynamic, moving around the battlefield, responding to or anticipating the helicopter’s route. As there is no research on the effect of the level of detail of threat modeling on route planning, the following research question will be investigated.

**RQ3:** How does the threat model’s level of detail impact the performance of the military route planning problem?

Finally, the decision-support tool will be applied to a comprehensive realistic mission. More particularly, an extension of Air Interdiction, namely Strike Coordination And Reconnaissance (SCAR) will be investigated. SCAR missions are defined as detecting and coordinating or performing an attack or reconnaissance on a target in a prespecified geographic zone [43,48]. The difference between SCAR missions from Air Interdiction
missions is that the exact location of the target is unknown beforehand. The target has to be searched and detected in a mission area. In practice, this extension of the problem is done by adding an objective of maximizing the search gain. Important to note is that the target is not necessarily static. It can move around the battlefield, responding to the location of the helicopter. Therefore, game-theoretic notions of searching are required. This natural extension of applying interactive MOO to planning the route for Air Interdiction missions gives rise to the following research question:

**RQ4:** How can an interactive optimization technique help determine the optimal helicopter route in searching for a target in a specified geographic zone (SCAR mission)?

In general, the decision support tool will serve as a basis for incorporating human-AI collaboration within the MDMP. In particular, it will serve as a basis for the Helicopter Warfare Center of the DHC to improve its mission planning using interactive optimization. Moreover, the active involvement of domain experts in designing interactive MOO techniques is a valuable contribution to the interactive MOO community. Note that the current mission selection can be changed based on feedback from the DHC. To answer all of the aforementioned research questions, a detailed research plan is needed.

### 3. Research Plan and Proposed Methodology

A detailed explanation of the five-stage research plan can be found in Figure 1.

**Stage 1: Defining the State-of-the-Art**  The first stage will consist of two literature studies integrated with ongoing extensive discussions with domain experts. The goal of the discussions is to get information on how route planning is done within the DHC. It will include, but will not be limited to, modeling the current planning process, listing relevant factors, and defining the mission scenario. To get this information both semi-structured interviews and attending multiple mission plannings are and will be done. Secondly, A scoping literature review of the military flight route planning problem will be performed. In this review, papers avoiding the threat zone will be excluded due to its unrealistic setting. This review will reveal criteria relevant to the military route planning problem and the current State-of-the-art modeling choices (RQ1). The result of this literature review
will be published in an internationally peer-reviewed journal. Simultaneously, a study on interactive MOO will be performed. In this review, the most appropriate interactive MOO method will be defined. The literature [30] classifies Interactive MOO methods on 1) when interaction is requested from the DM, 2) the type of preference information asked, 3) the preference model used to model the DM’s preferences, 4) and the search engine [30]. The exact modeling choices of these four points are to be done in collaboration and experimentation with domain experts. After all, it is of the utmost importance that the developed method is well-tailored to the domain experts. For example, the preference information has to be cognitively easy for the pilot to provide whilst being informative enough for the preference model to capture the DM’s preferences. Finally, an evaluation framework, based on the aforementioned desired properties, will be defined.

Stage 2: Developing an interactive MOO method

The route planning process is a process that involves a heterogeneous group of DMs making a collaborative decision. For example, intel officers, pilots, etc. are all included in the decision-making, necessitating group decision-making strategies. The interactive MOO method will need to be tailored perfectly to these domain experts, limiting simply reusing existing techniques. Therefore, based on the results of Stage 1, a novel interactive MOO method will be developed. Moreover, as the interface of interactive methods is crucial to the success of the method [29], a graphical user interface will be developed. The programming for both the interface and optimization method will be done using Python since this is mostly used by the MOO community (see for example, DESDEO [49] or pymoo [50]). Moreover, it will be done in an object-oriented programming fashion to ensure modularity of the code. This will enable easy extension of the model for Air Interdiction missions to SCAR missions.

Stage 3: First Experiment; Air Interdiction Mission

To answer RQ2, a user experiment that applies interactive MOO to Air Interdiction will be designed. The goals of this experiment are to validate the utility of human-AI collaboration within route planning and to further tailor the method to its users based on received feedback. The general idea is to compare manual route planning, a-priori, a-posterior, and interactive optimization methods with one another. More particularly, we will evaluate them on algorithmic and DM-focused performance. To measure these performance metrics, domain experts of the army will be used as DM in the project. The next stage aims to investigate the most important factor of military helicopter route planning, namely, threat modeling.

Stage 4: Threat modeling

To answer RQ3, different levels of detail of threat modeling will be compared to one another. This experiment aims to evaluate the effect of different levels of detail on the resulting route. Different test scenarios need to be defined, after which the resulting routes can be compared to each other.

Stage 5: Final Study; Interactive MOO for SCAR missions

To answer RQ4, a comprehensive mission scenario will be defined. This realistic SCAR mission will facilitate measuring the utility of the developed decision-support tool for complicated missions.

4. Acknowledgements

The supervisors for this Ph.D. project are prof. dr. ir. Roy Lindelauf, prof. dr. Herman Monsuur, prof. dr. ir. Mark Voskuil, and dr. Boris Cule. It is funded by the Dutch Defense Academy (NLDA) and the Data Science Center of Excellence.
References


Artificial Intelligence in Wearables - Challenges and Opportunities in Physical Therapy and Sports Training

Joanna SORYSZ

German Research Center for Artificial Intelligence (DFKI) and RPTU Kaiserslautern-Landau, Kaiserslautern, Germany
ORCiD ID: Joanna Sorysz https://orcid.org/0000-0002-9213-5468

Abstract.
Adherence to procedures and rules is essential in order to obtain the best results in medicine and sports. However, traditional clinical setups can induce stress in patients, hindering recovery. Meanwhile, advancements in activity recognition and monitoring technology have revolutionised the sports industry, yet systems suggesting exercises for performance improvement are sparse. At the same time, children training supervision is lacking comprehensive research altogether. In my research, I propose a project that aims to unify wearables solutions in physical therapy, sport training and children development. The key aspects of the research include the exploration of sensor modality fusion in order to obtain better results, body motion tracking, and physiological parameters recording. Planned experiments will focus on joint and torso movement mapping, integration with vital signs in order to perform real-life evaluations in cooperation with athletes, patients, coaches, and therapists.

Keywords. embedded intelligence, health care, sports, human activity recognition, human computer interaction

1. Context

More and more new sensors are created and improved. It is not a surprise anymore that the wear of mechanical parts can be measured in real-time or that during a football match, we can have statistics about the condition of the players. The wide selection of types and available modalities of detectors makes it possible to do a wide range of projects that would be otherwise impossible to finish. Diverse properties such as speed, exact position, sound intensity, applied force, pressure and many more are not a mystery anymore for researchers and innovators.

There are quite a few fields in which new technology proves to be of utmost value. Medicine and sports are the ones that are widely explored as they pose significant interest to many people. The first one is an essential part of our life without which more patients would be sick and die due to many various illnesses and accidents. It is only logical that a researchers are trying to improve tools and medication used for saving lives. Sport is an elemental piece of the nature of humans. Whether it is a professional career or a late
run for a bus or playing with friends, it is always present. It is not a surprise then that we want to know more about ourselves and the impact of sport on our bodies.

In medicine, new technologies and a better understanding of anatomy and physiology lead to the creation of more and more guidelines and norms. Adherence to procedures is necessary for patients’ safety. There are procedures in place when surgeries are performed, when the treatment plan is created, and when patients are guided through the first steps of recovery [9,23,36,33,3,25]. In every place in all hospitals in the world, rules are strictly followed and obeyed which allows for easy routine and followings of the recovery process. However, a sterile clinic is not an optimal environment for the healing process as it puts psychological stress on patients: a necessity of being in a strange room, surrounded by strangers. It is believed that patients can heal faster at home where they can relax [8,19]. On their way out of hospital patients are usually equipped either with a physical therapy plan or a recommendation of visiting physician. Nevertheless, due to various reasons such as high costs, long distance, need to return to work or lack of necessary time, many convalescents have to resign from following recovery plan which can result in wrongly healed injuries or movement limitations [27]. Existing research has mostly concentrated on interactions with older people, sport rehabilitation, etc. [13,15].

At the same time, a great number of children start professional sports training at a young age [32,22,37,11,35]. Their training plan is created by coaches based on their work experience. However, adjusting the workload during the growth phase is crucial for the correct development of the human body [34]. Unfortunately, it is not always possible in large training groups where the coach has to take care of all the athletes. Moreover, due to great differences in children’s bodies, and the way they change over time, it becomes a challenge to properly guide them in their sport career. Right now, there is hardly any research performed in the field of youth supervision, except for a few, singular works about specific sports or coping with various disabilities [38,26]. A solution that would allow for monitoring of performed exercises, predict fitness progress and recovery outcomes would help in both: rehabilitation and young athletes’ training.

On the other hand, the sports industry in recent years was enhanced by recently developed activity recognition and monitoring technology. Various technological solutions were used and tested to obtain the best results in case of movement tracking: IMU sensors, camera tracking etc. [28]. However, despite a great number of producers of specialised devices, there are only singular ones that would suggest types of exercises to perform to improve one’s physical form. This type of setup could be helpful for individual training and for coaches to supervise the athletes’ development.

In my dissertation, I want to concentrate on closing the existing gap between an available technology, and the needs of patients and athletes for remote training supervision. The most important questions of this research that need to be explored are movement tracking in all environments, recording and analysing physiological parameters and making predictions of fitness development and recovery progress. To investigate possible designs for such tools a series of experiments will be performed. Firstly the body movement recording test will be held in a sports environment which should provide information about the optimal number and placing spots of sensors. In a similar way, the physiological parameters will be explored with additional consultations with medical doctors and physiotherapists. To answer the last question experiments combining movement and body data will be performed in cooperation with experts to ensure correct comprehension of analysed data.
1.1. Related work

Recently, one of the most popular area of sport science is movement recognition. There is a huge amount of papers that describe usage of various types of sensors in order to track and map body movements [16,4]. One of the most popular technologies is inertial measurement unit (IMU) that with use of accelerometer, gyroscope and (usually) magnetometer is able to precisely determine a position of the sensor and through that - position of tracked place [4,7,2]. However, there are a few challenges related to this technology such as existence of a drift which requires calibration phase in order to work properly. Another difficulties is the needed number of sensors for proper mapping of the body movement: to fully record the state of the joint (bending angle) two sensors are required. Despite existing issues, a lot of works obtained satisfactory results that allow for activity recognition [5,6,1,13]. Many types of sports were tested and for each one different sensors positions were used on athletes bodies as well as on used equipment [4,16]. In some of these papers researchers concentrated not only on the body tracking but also on the determining factors coming into creating a good team, proper assessment of athletes by coaches, physical strain assessment, and more.

On the other hand the capacitive sensing technology is more and more experimented with. While there are a few technical approaches to this method (sensing the body self generated electrical field or creation electrical field in an electrode and measurements of body capacitive changes) in a human activity recognition field those sensors are used with conductive textiles [6,14]. This produces systems that are easy to mount onto the human body and can be soft and easy to maintain (washing etc.) [12,31]. However, one of the disadvantage of this technology is that the gathered data can provide only information about the relative movement of body parts in relation to themselves [10,31]. This aspect makes it hard to use capacitive sensors as a stand alone solution.

Physiological data such as pulse, breathing pattern and more can be obtained in many various ways. However, taking into account that the final solution might be worn by athletes or recovering patients the bio impedance sensors were chosen. The electrodes are using low amplitude, high frequency current on the body which allows for measurements of the body impedance which is dependant on the body tissue percentage [18,29]. This parameter is utilised, among other applications, in cardiac disease examinations [17], as well as in works concerning human body tissue percentage [20], medical condition diagnosis [24], and cybersecurity authentication problems [21,30].

2. Research Idea

Presented above related papers individually are very well covering small parts of a broad field which is human activity recognition in areas of sport supervision and physical therapy assistance. Nevertheless none of those works presents a solution that would cover a few areas or could be easily transferred into another domain of interest. A system like this could be an invaluable help for people that deal daily with patients and athletes. The possibility of quick adjustment of a training or rehabilitation plan to individual needs is crucial for the success in most cases.

However creation of a solution like that poses a few questions that are pivotal to answer while development process occurs:
1. What type of data physical therapists require in order to make informed decision on recovery plans for individual patients?

2. What data coaches require in order to make informed decision on training plans for individual athletes?

3. What body parts need to be tracked for algorithm to be able to detect type of performed activity?

4. What type of body parameters (e.g. pulse, oxygenation, breathing tempo) are required additionally to tracking body movements in order to determine the training effects that performed activity has on the muscles?

5. What joints movement are required to be tracked in order to determine the whole body posture?

6. What combination of various sensors technologies is the most effective in terms of mapping specific joints movements?

To answer those questions it is necessary to perform a series of experiments. Starting from the most simple ones to determine the best fusion of sensors in attempt to get the most accurate and descriptive data. Through research focusing on gathering information about training types, decision making reasons etc. Ending with work on the final combination of previously observed data and creation a functional, easily adapted system for sport and rehabilitation purposes.

Due to the nature of described experimentation course a few challenges arise:

1. Integration of sensors data from different modalities;
2. Choosing the smallest number of sensors that would allow for accurate activity tracking;
3. Getting the sufficient number of participants from various sport backgrounds for long term study;
4. Getting the sufficient number of participants from various stages of rehabilitation for experiments;
5. Connecting all sensors together in a well working system;
6. Ensuring acceptable levels of data and system security in an Internet of Things system;
7. Creating an easy to understand and comprehend user interface for people from various background and all ages.

To overcome them it is essential to firstly, perform a throughout background research about already finished experiments and about fusion of various modality sensors; second, to begin cooperation with universities and institutions that concentrate on sport and physical therapy science; third, to ensure that the newest cyber-security procedures are in place; and fourth, perform study about user interface design including surveys in target groups of society.

3. Methodology

In order to confirm the proper work of a system each step will be tested. Experiments will be designed in small steps to ensure that nothing is missed. Ground truth for most of the experiments will be data obtained from video recordings of the experiments as it is a state of the art method in human activity tracking. The angles of the joints for first
testing stages can be obtained with one of the few programs used for bio-mechanical measurements such as Kinovea. This part will be done in groups: first, hint joints (knee, elbow) second, ball and socket joint (hip, shoulder) third, the construct of a few types of joints (ankle, wrist), and in the end mapping of the torso. This order was designed to gradually raise the difficulty level of bio-mechanical description of the levels of freedom and movement equations. In experimentation in which it is impossible to collect sufficient video data, other methods, like motion capture with active markers, are considered. As the later phase is planned to examine strictly activity recognition in natural sport or rehabilitation environment manually created (by qualified personnel) labels will be utilised. For the last part, which is designed to be user interface, specially designed surveys will provide required feedback from testers. Classification problems will be tried with traditional machine learning methods as well as deep learning algorithms. Gathered data will be evaluated with a series of state of the art statistical parameters such as F1 score and consulted with experts in physiotherapy, and sport training fields.

4. Future works

Each part of the research is designed as a different project to maintain order and clarity. Experiments will start with tests of joint movement mapping for ankle, knee, hip, elbows and wrists. In the next part the testing of the torso mapping will be performed. As in physical therapy and some sports the spine position is essential to the performed movement it is necessary to be able to track this part carefully. In those stages one of the priorities, apart from the accuracy of measurements, will be to minimalize number of required sensor in order to keep the cost of final system as low as possible.

After those experiments will be finished, the evaluation of a necessary physiological parameters to first determine the state of the body during performed exercises and second, to predict how it will behave in the future during similar activity.

This knowledge combined will allow to perform tests with real athletes, coaches, patients and therapist. This is a significant part of the research as it will allow fine tuning of previously created models with the aim in providing the best possible outcomes that would be useful for exactly those groups of people. In this stage it will be crucial to also create the best practices of data handling to ensure an acceptable level of security.

References


Xin He, Zhihao Liu, Gengze Shen, Xiang He, Jionghong Liang, Yu Zhong, Tianlong Liang, Jie He, Yue Xin, Chi Zhang, Dongdong Ye, and Guofa Cai. Article microstructured capacitive sensor with broad detection range and long-term stability for human activity detection.


Joan D Penrod, Kenneth S Boockvar, Ann Litke, Jay Magaziner, Edward L Hannan, Ethan A Halm,


[38] Jing Yao and Yanze Li. Youth sports special skills’ training and evaluation system based on machine learning. 2022.
Integrating Minds and Machines: The Role of Digital Externalization in Evolving Cognitive Architectures

Gleb VZORIN a,1

Lomonosov Moscow State University

ORCiD ID: https://orcid.org/0000-0003-2034-8007

Abstract. The advancement of digital technologies has raised concerns about their potential impact on human memory and cognitive autonomy. This research proposal explores the phenomenon of digital externalization, where reliance on mnemonic and informational tools such as smartphones and AI assistants may lead to a decline in natural memory capabilities. We critically examine the existing literature, including Sparrow et al.’s seminal “Google-effect” study, and propose an alternative theoretical hypothesis that digital externalization might not be inherently negative but could potentially have positive effects on cognition. We present four research questions aimed at investigating various aspects of cognitive externalization, including its intentionality, mediation by AI, and impact on memory processing. We outline three experiments designed to address these questions, incorporating methodological improvements and theoretical insights. Preliminary results from Experiment 2 support the intentionality of the “Google-effect”, while ongoing work aims to refine our understanding through additional experiments. Finally, we discuss the broader implications of digital externalization for cognitive psychology and hybrid intelligence research and outline future directions for theoretical refinement and empirical investigation. This interdisciplinary inquiry underscores the need for a holistic understanding of digital externalization to inform the development of future hybrid intelligent systems that enhance human cognition while preserving individual autonomy.

Keywords. Digital externalization, cognitive offloading, hybrid intelligence, Google-effect

1. Introduction

There is a prevalent concern that the advancement of digital technologies, which serve to augment our cognitive processes, could potentially impair our cognitive functions and autonomy [1, p. 73]. Particularly in relation to human memory, the reliance on omnipresent mnemonic and informational tools such as smartphones, Google, and ChatGPT might lead to a decline in memory capabilities, inversely proportional to the enhancement of these technologies. This phenomenon, known in psychology as digital externalization [2], has yet to be thoroughly integrated into the discourse on hybrid intelligence, where an in-depth understanding of nuanced changes in cognitive architecture is essential.

One of the most referenced empirical investigations into this issue within the field of psychology is the experimental study conducted by Sparrow, Liu, and Wegner [3].

1 Corresponding Author: Gleb Vzorin, vzorin.g@gmail.com
Their research, which comprises four studies, indicates that when individuals are presented with challenging questions, they are automatically predisposed to think of computers. Moreover, it was observed that when people anticipate having future access to information, their ability to recall the information itself diminishes, while their ability to remember where the information can be accessed improves. This phenomenon was called the “Google-effect”.

Sparrow et al. also revealed that mnemonic externalization occurs irrespective of an individual's intention to remember information, suggesting that the decline in natural memory capacity might be an automatic response to the availability of external memory aids. This finding has potentially bleak implications for the future of hybrid intelligent humans. The inadvertent nature of cognitive offloading, especially when paired with increasingly sophisticated tools like ChatGPT, could significantly spoil human cognition then.

Efforts are underway to mitigate these possible concerning outcomes. Some studies focus on altering the current dynamics between humans and digital tools to foster more beneficial interactions, such as employing memory cues rather than complete content externalization [4]. Although these approaches show promise, it's uncertain whether major corporations will adopt them, as the trend towards increasing content externalization seems unstoppable. On the other hand, theoretical [5,6] and empirical research [7,8] suggests that the initially observed effects might not be consistently replicated, arguing that our cognitive functions, honed over millions of years, are robust against the transient impact of the digital age. Nonetheless, this line of inquiry cannot be deemed entirely successful, as this theoretical position is arguable [9], and subsequent studies have still managed to replicate the so-called "Google effect" [10].

We present an alternative perspective on digital externalization, suggesting that its effects might not be inherently negative but rather potentially positive. Our hypothesis posits the negative consequences of digital externalization are not due to a malfunctioning old memory mechanism but rather to an insufficiently evolved new mechanism. Therefore, we aim to explore the characteristics of this worsened memory and to determine the nature of the potential new mechanism.

The impression of memory degradation may arise from a viewpoint that sharply delineates between humans and computers. In anticipation of a future dominated by hybrid intelligence, which is poised to redefine human intelligence [1, p. 58], it becomes imperative to reassess our understanding of what constitutes a decline in memory capabilities. Memory impairment should not be judged by an individual's inability to execute tasks without external aids like Google or ChatGPT, but rather by their failure to adeptly manage these aids. The essence of human oversight lies in the ability to manipulate high-level concepts and chart the course of action. Memory is considered compromised only when an individual's grasp of these overarching concepts is insufficient for the flexible operation of devices. In this context, a well-functioning human “memory” is more about strategic thinking than conventional recall. It involves maintaining the core essence while offloading detailed information and derived meanings to external aids, thus freeing up cognitive resources for higher-level processing. However, traditional aids like notebooks or Google searches may fall short, as they lack the flexibility needed for effective task management, leading to mistakes perceived as memory decline. Conversely, generative AI technologies offer greater adaptability by enabling the on-the-fly reconstruction of meanings from the basic understandings retained in human memory, thus enhancing high-level cognitive functions and reasoning in future hybrid intelligence systems.
Should this hypothesis be true, it implies that even current cognitive offloading is not an automatic epiphenomenon (contrary to the findings of the original Sparrow study). Moreover, this selective offloading should not encompass all information; instead, the essential themes or core ideas should remain intact, or they might even be enhanced in human own memory.

2. Research questions

In accordance with our hypotheses, we propose four research questions that range from narrow empirical research questions to the broader theoretical framework of cognitive externalization in future hybrid intelligence systems:

1. Is there an inclination towards digital technology usage when individuals face complicated situations? This was initially observed in the Sparrow experiment with the emotional Stroop task, which was primed by the intention to 'google' challenging questions. However, subsequent attempts to replicate these findings were unsuccessful [11,12,13]. We speculate that this inconsistency may stem from the lack of a precise empirical framework in the original and replication studies.

2. Is the mnemonic 'Google effect' a result of automatic processes, or is it intentional and driven by specific purpose? Although the original study suggested that externalization does not depend on the participants' intention, we propose that this might be due to the participants implicitly setting a goal to remember information in all experimental conditions, even those that explicitly instructed them not to do so.

3. How does the effect of cognitive externalization differ when mediated by AI as opposed to search engines like Google? We hypothesize that with AI mediation, participants will be more likely to retain and focus on general concepts while offloading less critical details.

4. What is the metatheoretical foundation for the cognitive architecture of future hybrid intelligence systems? Our research aims to conclude by proposing potential mechanisms of digital memory integration, drawing from psychological theories.

3. Method

In Experiment 1, we aim to resolve a possible methodological limitation identified in the first Sparrow experiment, which we believe may have contributed to the failure of three subsequent replication attempts. Our concern is that the original framework, which utilized the Emotional Stroop task, lacked sufficient precision for detecting the effect in question. This may be due to the low ecological validity of using text labels (not images) as stimuli and due to the lack of semantic conflict with the color of the text. To address this, we propose employing a more suitable framework that allows for higher accuracy and the use of images as stimuli – the antisaccade task. In this experiment, participants will be presented with lists of trivia questions, categorized as either difficult or easy, followed by an antisaccade task featuring stimuli associated with the internet (e.g., the Google logo) and non-internet related stimuli (e.g., the Nike logo). We hypothesize that exposure to internet-related stimuli, following difficult trivia questions, will result in increased latency and a higher error rate in comparison to non-internet related stimuli. This effect is expected to be absent or significantly reduced after easy trivia questions.
Experiment 2 is designed to further investigate the mnemonic "Google-effect" by incorporating a more nuanced approach to controlling participant intent. According to our theoretical hypothesis, digital externalization is an intentional feature, rather than an unintentional "bug" of human memory. This is supported by the success of a recent replication study [10], which managed to effectively influence participant goals by demonstrating the reliability of the experimental procedure. The aim of our Experiment 2 is to address a potential oversight in the original studies, where the effect of externalization appeared to be independent of the experimenter-defined objectives, possibly due to an intrinsic motivation among participants to remember information regardless of the experimental conditions.

To subtly influence participants' intrinsic motivations, the experiment will employ a detective-like game that allows for the combination of three different intents (to remember, not to remember, and no explicit goal) with three actions (save, delete, and no action). This setup creates a 3x3 matrix of conditions, each backed by a compelling rationale to ensure participant engagement and understanding.

The assessment of memory will be conducted one week after the initial phase, integrated into a non-memory-related detective task. Participants will be required to select from a set of previously encountered stimuli to solve a detective puzzle. This approach hypothesizes that the "Google-effect," characterized by a diminished retention of saved information, will be most apparent when participants are explicitly instructed to remember or when no specific goal is provided, mirroring the conditions of the original study. However, in scenarios where participants lack a genuine intent to remember, an automatic saving process is expected to occur, facilitated by the use of a recognition-based test rather than recall. This saving process is anticipated to be uniform across different operation actions, indicating that the intentional aspect of memory externalization plays a crucial role in the "Google-effect."

Experiment 3 is aimed at proving that digital externalization is not merely an intentional offloading of all possible memory content onto an external device. The underlying premise is that digital externalization is a deliberate, purpose-driven process aimed at sparing cognitive resources for higher-level memory tasks by transferring lower-level information processing to technological aids. This assumption is partially supported by the experimental study conducted by Lu et al. [14], which concluded that externalization does not negatively affect gist memory. However, this conclusion requires further examination in light of our hypothesis concerning AI mediation.

In Experiment 3, participants will be exposed to various text stimuli and then instructed to remember them. Participants have to perform three types of operations for managing this information: deleting it, saving it to notes, or saving it to ChatGPT. This design allows for the comparison of different forms of digital externalization.

A week later, participants will be tested on their recollection of the stimuli (without any external devices) through questions designed around the modified Levels of Processing framework [15]. This test will include four categories of questions: perceptual (e.g., the color of the text), factual (e.g., details like the color of the curtains in a scene), semantic (e.g., the actions of a character), and the overarching gist or themes that were not explicitly stated in the text (e.g., the character's underlying motives).

The hypothesis posits that participants will offload less meaning in the AI ("ChatGPT") condition compared to "Google" condition. This expectation is based on the assumption that AI, like ChatGPT, offers a more interactive and sophisticated form of externalization, potentially enabling users to retain higher-level conceptual information more effectively than when using more straightforward search tools. The
results of this experiment could provide valuable insights into the strategic choices individuals make when deciding what information to externalize and what to retain, highlighting the cognitive strategies employed to manage the vast amounts of information encountered in the digital age.

Addressing the broad implications of digital externalization within future hybrid intelligent systems indeed requires a robust theoretical framework alongside empirical studies. Integrating concepts from the theories of "cognitive gadgets" by Cecilia Heyes [16] and the "higher mental functions" by Lev Vygotsky [17] can provide a multidimensional understanding of the potential evolutionary psychological mechanisms underpinning the memory architecture of future hybrid intelligence.

The concept of "cognitive gadgets," as proposed by Heyes, suggests that cognitive skills are primarily cultural products, learned and refined through social interaction rather than being entirely innate. This perspective can be particularly relevant when considering digital externalization, as it frames the integration of digital tools into cognitive processes as a form of cultural evolution, where tools like AI and digital devices become part of our cognitive toolkit through social learning and cultural transmission.

On the other hand, Vygotsky's theory of "higher mental functions" emphasizes the role of social environment and cultural tools in cognitive development, arguing that complex cognitive processes are mediated by symbols and tools that are provided by the culture. This theory supports the idea that external cognitive aids (like digital tools) can become integral components of human cognitive processes, transforming basic, innate mental functions into more sophisticated and culturally augmented ones.

By combining these perspectives, a theoretical model can be developed that envisions future hybrid intelligence systems as the next step in the cultural and cognitive evolution of humanity. This model would propose that through the continuous integration of digital tools and AI into our cognitive practices, we are not merely offloading memory tasks but are fundamentally reconfiguring the architecture of human memory itself. This reconfiguration could lead to the emergence of new forms of cognitive processes that are inherently hybrid, leveraging both innate human capacities and the vast capabilities of artificial intelligence.

Such a model can also be used in exploration of the implications of these transformations for individual identity, social interaction, and collective knowledge construction, considering how the boundaries between human cognition and artificial intelligence may blur, leading to new challenges and opportunities in understanding what it means to be intelligent in a digitally augmented world.

4. Preliminary results

To current date there are only results of Experiment 2 available (N=69). The obtained results generally replicated the original mnemonic "Google-effect": under conditions of an explicit "remember" goal and in the absence of a clearly stated mnemonic goal, there was an advantage in choosing stimulus material for which the operation of "deletion" from an external medium was performed, compared to the "save" operation and skipping the operation.

A result corresponding to our empirical hypothesis was also obtained. In the condition where the game's narrative explicitly set the material's uselessness and aimed to "not remember" it, regardless of the operation performed on such material, an equal probability of recognizing it in the second series was observed. Notably, the equal
probability result for all three operations is not explained by the absence of imprinting due to the early-stage filtering of information as irrelevant. On the contrary, the accumulated frequency of errors (choice of distractors) in this condition was significantly lower than the probability of choosing preceding stimuli and did not differ for all three mnemonic goals, indicating the preservation of mnemonic traces. Thus, subjects imprinted the stimulus material at the level of reproduction in the "not remember" goal condition, however, this material was not differentiated or prioritized depending on the fact of its preservation on an external medium or deletion.

Preliminary results for our ongoing research are currently being compiled, with Experiment 1 actively in its conducting phase. This experiment, which revisits and refines the methodology of the original Sparrow experiment through the implementation of the antisaccade task, is expected to provide nuanced insights into the methodological challenges previously encountered in replications. We anticipate completing the participant acquisition for Experiment 1 within the next three months.

Experiment 3, on the other hand, is in its planning stage. Our objective with this experiment is to explore the intentional aspects of digital externalization and its effects on various levels of memory processing. By employing a detective-like game to modulate participants' intrinsic goals, we aim to delve deeper into how digital tools, particularly AI, influence memory retention and offloading. The planning phase for this experiment is projected to last one month, with an additional four months allocated for participant acquisition.

The projected timeline for these steps ensures a structured approach to our research, allowing for meticulous data collection and analysis, which will contribute to the broader discourse on the evolution of human cognition in the digital age.

5. Discussion and Future work

Our intermediary conclusion posits that adopting a broader perspective on digital externalization not only facilitates the convergence of cognitive psychology and hybrid intelligence research but also presents opportunities to address prevailing challenges within both domains, particularly concerning human memory. In the realm of cognitive psychology, digital externalization is mired in controversies, some of which have been delineated in the preceding sections of our paper. By envisaging memory not as a static entity but as a dynamic construct that evolves through the integration of new mechanisms within future human-AI systems, we can empirically address these contentious issues.

The primary focus of our future work will be on the successful execution of Experiments 1 and 3. In addition to the empirical work, a significant portion of our future efforts will be dedicated to refining and articulating our theoretical model. This model aims to bridge cognitive psychology and hybrid intelligence research, providing a comprehensive framework that accounts for the evolving nature of human memory in the context of increasing digital integration. It is imperative that this model is presented in a clear and accessible manner, enabling both scholars and practitioners across disciplines to grasp its implications and applications. By elucidating the theoretical underpinnings of our research, we hope to contribute to a more nuanced understanding of digital externalization and its impact on cognitive processes, thereby informing the development of future hybrid intelligent systems that enhance human cognitive capabilities while preserving individual autonomy and agency.
References


Hypothesis Updating by Combining Knowledge Graphs and Argumentation

Giacomo ZAMPROGNO 1
Vrije Universiteit Amsterdam

Abstract. This project explores the applicability of computational argumentation and knowledge graph methods for the validation and updating of scientific hypotheses. Starting from a game setting example, we suggest that such framework could be extended to the scientific discovery domain as a long-term plan.

Keywords. Computational Argumentation, Hypothesis Representation, Hypothesis Evolution, Scientific Knowledge Graphs

1. Introduction

Motivation: developing scientific assistants. In recent years, there has been an extremely fast increase in the volume of scientific publication: for example, science and engineering publications increased from about 2 millions in 2010 to 3.3 millions in 2022 [1]. As a consequence, delving into the amount of literature can often feel an intractable task. In this sense, it feels natural to explore the possibilities for the creation of scientific assistants able to help with sifting through the literature and validating/evolving the hypothesis proposed by the researcher. Following the recent rise of Large Language Models (LLMs), various attempts of assistants have been studied [2,3] but, despite their clear potentialities, they share some of the common explainability pitfalls of deep learning approaches (hallucinations, bias and retrieval of non existent sources are often mentioned as possible concerns [4]).

Example Scenario. Consider a scenario where a team of researchers is trying to develop new treatments for asthma. Starting from the findings of paper A, they consider studying the introduction D-vitamin in the therapy as paper A links it to reduced symptoms severity and control. They then input the research question and the related literature supporting the idea to an assistant. The assistant consults a paper [5] in its knowledge base, where a systematic review on diabetes and D-vitamin correlation is performed, with milder conclusions. The agent suggests lack of strong evidence for a positive interaction. An assistant could help the researchers to further develop or revisit their hypothesis: in

1Supervisors: Ilaria Tiddi, Vrije Universiteit Amsterdam, NL; Bart Verheij, Rijksuniversiteit Groningen, NL.

This work is partially supported by the Hybrid Intelligence programme (https://www.hybrid-intelligence-centre.nl/), funded by a 10 year Zwaartekracht grant from the Dutch Ministry of Education, Culture and Science.
this simple scenario, the system could ‘argue’ against the original hypothesis, while the researchers would be given the option of providing additional supporting literature or updating their hypothesis after the attack, iterating the process until they are satisfied with the new revised hypothesis.

Towards an hybrid intelligence solution. Various stakeholders [6,7] have pushed for a shift of the publishing paradigm towards a more standardized and machine-readable approaches, as more and more research effort is put into the creation of annotated databases of scientific research. In this context, new opportunities and challenges appear, among which the possibility of automatically validating a given novel hypothesis against the machine-readable resources encoded in the form of Scientific Knowledge Graphs (SKG).

While SKGs provide structured knowledge for a scientific assistant, they might not be enough to foster the development of Hybrid Intelligence [8] systems, where humans and agents are intended enhance each others’ capabilities toward achieving a shared common goal in an explainable, collaborative, responsible setting. Computational argumentation could then represent an inherently explainable approach to bridge the gap: considering that the scientific discourse is argumentative by nature, one could thus imagine a scientific assistant that opens a dialogue with the researcher, highlighting the possible weaknesses (attacks) and strengths (support) of the proposed hypotheses and helping their validation and revision via an iterative process.

2. Related work

This section will provide an overview of existing literature on the two more technical aspects of this project: machine readable research and computational argumentation. Due to space limitations, we will leave a detailed discussion on how to integrate those techniques into hybrid intelligence systems for further work, as we expect it to require an in-depth exploration.

Machine-readable research: datasets and models. In the scope of providing machine readable scientific knowledge, Knowledge Graphs have been employed [9]. Many SKGs are already available, but they are usually limited to a single area of expertise: to mention a few examples, mentioned in [10], Papers with Code2 for machine learning, MetaLab3 for cognitive sciences and coda4 for social sciences.

There also exist attempts to provide a systematic collection of research works from different domains, populating graph data model from annotated papers: the Open Research Knowledge Graphs (ORKG) [10] aims to collect an extensive amount of papers from different areas, and present them in accordance with FAIR (Findability, Accessibility, Interoperability, Reusability) guidelines [11]. PubGraph [12] is a large scale knowledge graph, which maps the metadata of the OpenAlex [13] to Wikidata [14] properties, further integrated with data from the Semantic Scholar Academic Graph (S2AG) [15], and it additionally provides large-scale benchmarks for knowledge graph completion tasks. OpenAIRE [16] is an European project that aims to provide a Graph Data Model to ‘foster, support and monitor Open Science scholarly communication in Europe’; be-

2https://paperswithcode.com/about
3https://langcog.github.io/metalab/
4https://cooperationdatabank.org/about-coda/
Beyond the model, an actual research graph is provided, with a focus on Open Science and highlighting links between scientific resources and findings.

Other options relate top-down approaches, defining generic models to represent scientific claims, their provenance and their evolution. For example, nanopublications [17] define a minimal model to encode independent scientific publications into Knowledge Graphs and reference them; a nanopublication consists of an assertion, its provenance and its related publication info (metadata), which could offer an initial bridging tool for available data (SKGs) and novel hypotheses. Similarly, the DISK framework [18] defines a model to keep track of hypothesis revision and updating in the context of automated discovery. It is specifically integrated as an open source framework for scientific knowledge discovery, and as we aim to focus on hypothesis updating, it could provide insights and functionality for tracking the evolution during the interaction.

Computational Argumentation. Computational argumentation stems from non-monotonic logic and defeasible reasoning and it is generally concerned with the acceptability of sets of statements and arguments that might be incompatible, or alternatively provide support, to each other.

A large number of modern argumentation frameworks is originated from the introduction of Abstract Argumentation Frameworks (AFs) [19]. In this seminal work, the author ‘abstracts’ from the content of the arguments per se and focuses on analyzing the relations between arguments. Specifically, only attacks are considered, and intuitively an argument can be accepted if either it has no attackers, or if additional arguments are (recursively) accepted such that they attack all the original argument’s attackers.

[19] also introduces various semantics defining ‘rational’ ways to justify acceptability of sets of arguments. As an example, the grounded extension includes the arguments that can be accepted without disagreement (i.e. representing the ‘common ground’ in a debate), while an admissible extension is a set of arguments where each element is admissible, possibly different from the grounded counterpart as multiple admissible extensions could exist, contrasting with each other (e.g. given arguments A and B attacking each other, both the single element sets \{A\} and \{B\} are admissible, while only the empty set \{} is grounded).

Bipolar Argumentation Frameworks [20], and Quantitative Bipolar Argumentation Frameworks (QBAF) [21] further extend AFs by respectively introducing the support relations and a quantitative weighting to attacks and supports. In QBAFs, semantics change from value-based, where arguments are accepted or rejected, to quantitative: each argument initially starts with a given strength, and semantics defines how such values are updated depending on attackers and supporters.

An alternative approach to include quantitative information in AFs is proposed by [22] with Weighted Argumentation Frameworks (WAF), which are value-based and extend original AFs by defining the concept of a inconsistency budget, where attacks can be disregarded until their combined strength reaches a given threshold.

In the context of this project, we expect to not only consider the relations among arguments, but the content of the argument themselves, and how (scientific) hypotheses can be mapped into argument structures. The work by Toulmin [23] could be seen as a promising starting point to the aspects of hypotheses representation. He proposes a standardized high-level model of arguments, which at its core defines the concepts of claim, data (evidence) and warrant (the explanation that links data and claim). The model and its extensions could be adapted to represent and compare scientific claims.
Alternatively, Argumentation Schemes [24] (AS) also define templates of arguments, often following a structure similar to *modus ponens*. Argumentation Schemes try to provide a compendium of common argument patterns, together with a list of *critical questions*, which if true could be used as defeaters. They could thus provide a model for the discourse on scientific hypotheses verification: finding an appropriate AS for mapping the claims, the critical questions could be used to guide the search in the knowledge base for possible attacks. Argument schemes are also formalized in the Argument Web [25], which is based on the Argument Interchange Format (AIF) [26]. These tools aim to provide a standardized, shareable and interoperable ecosystem of various computational argumentation resources and approaches in the context of the Semantic Web [27].

3. Research Questions and Challenges

The wider research question we aim to tackle with this proposal is *to what extent a combination of scientific knowledge graphs and argumentation frameworks can provide assistive functionalities for hypothesis verification and updating.*

Specifically, we propose the following sub-questions as steps towards the core one:

- **Can we re-utilize existing formalization and models for representing scientific knowledge and hypotheses?**
  In the context of FAIR science, it is fundamental consider reusability of models and thus preference should be given to extending what resources already exist and are consolidated. As a consequence we aim to adapt and extend existing models and verify if they are sufficient to capture the information needed to compare scientific claims.

- **Which argumentation framework already exist that can provide updating functionality for hypotheses?**
  Multiple argumentation frameworks have been proposed in the literature, often with different expressivity and use-cases. We plan to investigate which of the available formalization options can be used for the described research problem, and which trade-offs between their complexity and expressivity are involved in the task of comparing hypotheses.

- **Are the semantics related to such frameworks suited to verify and update hypotheses from SKGs?**
  Each argumentation framework offers different types of semantics for the acceptability of arguments. While the more common concept of grounded and admissible semantics can already represent valuable evaluation tools, it might be the case that more on-point semantics need to be suggested for formalising the condition for accepting a proposed hypothesis and its justifications.

- **How can a scientific assistant support researchers by aligning with the principles of Hybrid Intelligence?**
  Aiming to achieve an hybrid setting, a scientific assistant should be designed to enhance and support the capabilities of the human users. Being able to sustain an argumentative dialogue could be a first requirement, but we also plan to explore what other collaborative techniques could foster positive interaction.
4. Approach and evaluation

Figure 1 provides a high level visualization of the pipeline we intend to implement. Chosen an appropriate scenario, we will create and populate a dataset in the shape of a knowledge graph (0), which we will use to generate arguments to test the hypotheses. In the pipeline we propose, the user’s hypotheses (1) will be mapped into a common representation (2): initially, as we intend to try to reuse and adapt available resources, we plan to study the possibility of using Toulmin’s model and nanopublications for such representation. The represented hypothesis and the information contained in the KG will then be aligned (3), and arguments in support or contrasting the hypothesis will be identified, thus creating an argumentation graph. In the initial phase, we will employ simple techniques tailored on the toy ontology akin to pattern matching. As argument schemes have already been widely treated in the literature, the first iteration of the pipeline development will could then finding schemes that can map relevant reasoning strategies and match them to the information available in the graph. In the future, we also expect to shift towards more complex techniques as we tackle the scientific assistant topic, possibly using machine learning-based methods like graph embeddings or graph neural networks. We intend to study the applicability of increasingly complex argumentation frameworks for validating the hypotheses against the available knowledge base (4), ideally QBAs or WAs would represent the most expressive options as they allow quantitative information, but in the first phases we will first test the potentiality of simpler frameworks as AFs, which could nonetheless provide baseline functions in detecting faulty hypotheses.

The results of the argumentation will be an ‘argumented hypothesis’, containing some criticism or some additional suggestions for strengthening the claim. Iteratively, the user will be able to provide a refined version of the hypothesis, possibly answering to the attacks, until a satisfying result is reached. Ideally, the system will also keep track of the various updates, for which we initially intend to adapt the DISK framework.

We plan to develop the project incrementally, starting from a controlled ‘toy’ context and later applying the results to the more generic scientific assistant area. At various steps of the iteration, we plan to evaluate our results via user studies. Additionally, performance and ease-of-use might be further evaluated by comparison with deep-learning based (scientific) assistant models.
5. Current results and Conclusion

Current results. The context of choice for the toy setting is the social deduction board game ‘Avalon’, in which players constantly create hypotheses regarding the hidden roles of other players given shared evidence. There are multiple reasons for which we chose the game as a starting point. First, its nature reflects a core part of the problem at hand: as a social deduction game, players are required to formulate hypotheses about the hidden role of other players by using imperfect and limited evidence. Second, the search space is restricted by the ruleset, creating a more controlled use-case, which can be more easily analyzed compared to directly tackling the task of scientific knowledge discovery. Third, a large corpus of open data is available from the online platform ProAvalon\(^5\) which will allow us to directly approach the development of a prototype, bypassing the need to develop experiments and data collection for the initial phase.

In the next steps we plan to develop a ‘proof-of-concept’ system. While we have full control over the Avalon dataset, scientific resources contained in SKGs often vary in representation and completeness of information. This will require us to explore ways of mapping such variety of representation into the chosen argument models, overhead we can temporarily ignore in Avalon where the arguments can be hand-crafted from game knowledge.

At present, we built a Knowledge Graph based on the Avalon dataset for which we designed and populated an owl ontology. It currently contains 19 classes and 17 properties, and maps data from the online matches and from the rules of the game, which we will employ for the argumentation process. It has a total of 2700 axioms from the mapping of the first 10 game entries of the dataset, but it can be easily extended to a larger scale once the rest of the pipeline is developed.

As following steps, we aim to analyze which argumentation schemes can be used to map common reasoning patterns in the game, and try to represent role hypotheses and their justifications via such schemes, verifying and updating using critical questions.

For the user-based evaluation, the administrators of the ProAvalon platform have shown interest in possibly allowing the integration of bots as available game options, which could provide preliminary results regarding the applicability of the framework.

Conclusion. In this work, we presented a doctoral research proposal for combining knowledge graph and computational argumentation technologies for hypothesis validation and updating. After reviewing the main resources that we intend to employ for the project, we described the proposed pipeline and presented the current results of the project.

References


\(^5\)https://proavalon.com/statistics


Posters/Demos
This page intentionally left blank
Cognitive Network Science Unveils Affective Bias in GPT Models Mirroring Math Anxiety in High-School Students

Katherine ABRAMSKI a,1, Salvatore CITRARO b, Luigi LOMBARDI c, Giulio ROSSETTI b and Massimo STELLA c

a Dept. of Computer Science, University of Pisa, Italy
b Institute of Information Science and Technologies, CNR, Italy
c Dept. of Psychology and Cognitive Science, University of Trento, Italy

The introduction of Large Language Models (LLMs) has taken the world by storm, and society’s reaction has been anything but unanimous, ranging from humorous amusement to catastrophic fear. Among the most prominent LLMs are OpenAI’s GPT-3, GPT-3.5, and GPT-4. GPT-3 and GPT-4 are powerful and flexible models that can be fine-tuned to perform a wide variety of natural language processing tasks, while GPT-3.5 turbo is a variant of the other two, specifically designed to perform well in conversational contexts. All three belong to the family of generative pre-trained transformer (GPT) models that are trained on massive amounts of textual data to learn patterns and relationships in text. While these models have proven to be incredibly useful tools for everyday tasks such as composing emails, writing essays, debugging code, and answering questions, they have been shown to demonstrate harmful biases similar to the ones that humans possess. Biases in LLMs are misrepresentations and distortions of reality that result in favouring certain groups or ideas, perpetuating stereotypes, or making incorrect assumptions [1]. While these biases can be influenced by many factors, they largely originate from implicit biases in the massive text corpora on which the models are trained. Thus, the output produced by LLMs inevitably reflects stereotypes and inequalities prevalent in society. This is problematic since exposure through interaction with LLMs could lead to perpetuating existing stereotypes and even the creation of new ones [2,1]. Therefore, it is ever more important to understand the behavior and risks of these models. This challenge requires developing new benchmarks and methods for quantifying affective and semantic bias, keeping in mind that LLMs act as psycho-social mirrors that reflect the views and tendencies that are prevalent in society. One such tendency that has harmful negative effects is the global phenomenon of anxiety toward math and STEM subjects. Just as negative biases towards math and STEM are absorbed by children from their teachers and parents, LLMs acquire such negative biases from their training data. Understanding these biases in LLMs is essential, since at the societal level, math anxiety may deter capable students from pursuing careers

1 Corresponding Author: Katherine Abramski, katherine.abramski@phd.unipi.it.
in STEM, especially females. In this work [3], we investigate biases produced by LLMs, specifically GPT-3, GPT-3.5, and GPT-4, regarding their perception of academic disciplines, particularly math, science, and other STEM fields. To accomplish this, we apply behavioral forma mentis networks (BFMNs) as a method of investigation. BFMNs are a type of cognitive network model that capture how concepts are perceived by individuals or groups by building a network of conceptual associations [4]. To build such a network, we gather data obtained by probing the three LLMs in a language generation task that has previously been applied to humans. We repeatedly asked the LLMs to produce associative responses to the various cue words related to academic disciplines (e.g. math, science). From these cues and associated responses, we built associative networks such that cues were linked to all of their responses. Furthermore, we asked LLMs to provide sentiment ratings (positive, negative, or neutral) for all cues and provided responses. These sentiment ratings were used to enrich the networks with node features. We thus obtained feature-rich behavioral forma mentis networks representing conceptual knowledge related to the cues. To better understand this conceptual knowledge, we applied semantic frame analysis to investigate the biases that emerge within these networks with respect to the cues. Our findings indicate that LLMs have negative perceptions of math and STEM fields, with the most negative biases toward math compared to other academic disciplines. These findings mirror the negative attitudes of high school students from previous work [4]. Despite overall negative perceptions, we observe significant differences across OpenAI’s models: newer versions (i.e. GPT-4) produce semantically richer responses with more emotionally polarized perceptions and fewer negative associations compared to older versions and high school students. These findings suggest that advances in the architecture of LLMs may lead to increasingly less biased models that could even perhaps someday aid in reducing harmful stereotypes in society rather than perpetuating them.

Figure 1. Sentiment enriched semantic frames for math produced by GPT-3, GPT-3.5, and GPT-4. GPT-3.5 and GPT-4 produced much richer semantic frames compared to GPT-3, and GPT-4 produced a significantly more positive semantic frame compared to GPT-3 and GPT-3.5.
References

Why Do You Think This Person Is Introverted?

Towards Conversational Argumentative Explainability

Alexander Berman

Abstract. This paper presents a system that supports conversational argumentative explainability between AI and human.

Keywords. human-AI collaboration, hybrid human-AI intelligence, conversational explainability, argumentation theory, explainable AI

1. Introduction

Previous work has shown that users of AI-based decision-support systems prefer conversational forms of explanations, i.e. interfaces that enable users to request and obtain explanations for model predictions in natural-language dialogues between user and system [1]. However, issues regarding how to design, implement and evaluate such interfaces remain relatively unexplored (exceptions include [2,3,4]). One of the challenges that requires more attention concerns feature contributions, i.e. how an AI prediction is underpinned by particular features (factors). In everyday conversation, relations between features and judgements are often expressed as rhetorical arguments, as in the utterance “Let’s walk along Walnut Street. It’s shorter.” [5, emphasis added], where a route’s short distance is presented as support for its goodness. Typical methods for explainable AI, such as LIME [6] and SHAP [7], only convey how much a feature contributes to a prediction (i.e. feature importance), not how they contribute. In natural language, this corresponds to merely referring to features by their name: “Based on distance, I suggest that we walk along Walnut Street.” (cf. [2,8]). This communicative strategy might be successful in cases where the explainee can easily infer the nature of the contribution. For example, if the explainee already knows that the route via Walnut street is shorter than other options at hand, merely referring to distance may suffice. But if the goal of an AI-based decision-support system is to facilitate hybrid human-AI intelligence [9] – including the possibility to make better decisions with the help of AI – explanations cannot assume that an AI always “reasons” in similar ways as humans. In line with this, one of the aims of the prototype described in this paper is to show how feature contributions in the form of more explicit arguments can be supported in a dialogue interface between a human and a statistical model.

1Corresponding Author: Alexander Berman, alexander.berman@gu.se.
2. Prototype

We demonstrate **MindTone**\(^2\), a browser-based game featuring conversational argumentative explainability between AI and human. The task of the game is to estimate whether persons are extraverted or introverted based on their music preferences. In each round of the game, the player is shown specific tracks that a person has listened to frequently as well as audio statistics of music heard by the person, such as loudness and energy. The player is assisted by a chatbot that predicts whether the person is extraverted or introverted using a logistic regression model trained on audio statistics (mean values for loudness etc.) as independent variables and degree of extraversion as dependent variable [10]. The dialogue manager is implemented in a rule-based information-state update approach [11] inspired by conversation-oriented semantics [12] and argumentation theory [13,5]. It supports various dialogue strategies including feature contribution arguments, as exemplified by the following dialogue between user (U) and assistant (A):

A: I think this person is introverted.
U: Why?
A: The person likes music with high energy.
U: I don’t understand
A: Statistically, people that like music with high energy tend to be more introverted.
U: OK, I see. Why do you think this person likes music with high energy?
A: Music heard by the person has a higher average score for energy than music in general.
U: OK, and does the the fact that the person likes music that is not danceable support the assessment that the person is introverted?
A: Yes.

For natural-language understanding, the system uses GPT-4 [14] to parse user utterances into a formal language for dialogue moves, while templates are used for natural-language generation.

3. Future Work

In the future, it would be useful to empirically study how conversational strategies such as feature contribution arguments affect human-AI team performance, e.g. compared to a similar interface that only supports feature importance. Extending the prototype to support model-agnostic explanations (rather than only logistic regression) could also be an interesting avenue for further work.

Acknowledgements

This work was supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

---

\(^2\)Live demo: https://github.com/alex-berman/argumentative-explainability
References


Pei-Yu CHEN\textsuperscript{a,1}, Selene BAEZ SANTAMARIA\textsuperscript{b}, Maaike H. T. DE BOER\textsuperscript{c}, Floris DEN HENGST\textsuperscript{b}, Bart A. KAMPHORST\textsuperscript{d}, Quirine SMIT\textsuperscript{c}, Shihan WANG\textsuperscript{e}, and Johanna WOLFF\textsuperscript{f}

\textsuperscript{a}Delft University of Technology, The Netherlands
\textsuperscript{b}Vrije Universiteit Amsterdam, The Netherlands
\textsuperscript{c}TNO, The Netherlands
\textsuperscript{d}Wageningen University, The Netherlands
\textsuperscript{e}Utrecht University, The Netherlands
\textsuperscript{f}Twente University, The Netherlands

Abstract. Behavior change support systems need to take into account individual needs and preferences to provide appropriate support. In this demonstration, we illustrate how this might be achieved through the explicit modeling of user characteristics within knowledge graphs (KG), captured in a dialogue between the system and the user. We demonstrate how up-to-date information enables reasoning for providing personalized support.

Keywords. Dialogue, Knowledge graphs, Reasoning, Behavior support system

1. Introduction

For behavior change support systems to offer adequate support, they should be able to adapt to the diverse and evolving nature of the users in unforeseen circumstances \cite{1}. One way to adapt a system is by implicitly learning users’ preferences in different circumstances from behavior data. However, behavioral data reflects people’s past behavior rather than their future desired behavior. Capturing the latter is particularly important for systems intended to support a user in changing their behaviors. In this demo, we propose a complementary approach that accurately and explicitly represents important domain-specific information (domain KG) and user-specific information such as context and its influences on norms and values (user KG). Besides the ability to store dynamic and static knowledge, KGs offer transparency and explainability, as the system’s reasoning process becomes explicit \cite{2}. This user model must be updated at run-time to capture, for example, the changes in context \cite{3, cf.} through direct dialogue. This dialogue be-

\footnote{1Corresponding Author: Pei-Yu Chen, p.y.chen@tudelft.nl. Authors are listed alphabetically by last name, except for the first author.}
The system integrates five key components, as shown in Figure 1. Below we elaborate on the technical aspects of each component and illustrate their roles in creating a comprehensive support system that adapts to individual user needs and contexts.

1 **Dialogue** A dialogue component engages users in “Alignment Dialogue” [6], designed to gather information crucial to ensuring the support provided is in alignment with the users. Given the complexity inherent in such dialogues, a rule-based approach is employed in this demonstration to manage and guide the interactions effectively.

2 **Information Extraction** This component transforms text (e.g., “I love walking”) into a subject-predicate-object structure (e.g., Pedro-like-walking). We use RDF triples and named graphs to recursively combine knowledge units into complex structures capable of expressing the content, form, and context of a dialogue [5].

3 **User KG** The User KG organizes RDF triples from the Information Extraction Component. The User KG uses an OWL-based Ontology that includes concepts to represent user contexts, preferences, values, and others. The User KG contains the user’s health data such as blood sugar and weight, as well as the user’s values, preferences, and other important factors. This information is used to make personalized recommendations.

4 **Domain KG** The Domain KG contains medical knowledge about Diabetes and treatment options, e.g. which treatment types work best for which kind of user.

5 **Reasoning Engine** This component determines the required intervention type, based on the user’s health data from the User KG and medical information from the Domain KG.
KG. The system then uses the user’s preferences, values, and context from the User KG to select the most suitable action within this intervention type.

An interactive prototype\(^2\) of the support system will be available at the conference. Attendees will be able to converse with the system, see in real time how it updates its knowledge graphs, and examine the reasoning outcomes that inform the user dialogue. This prototype aims to showcase the system’s potential for supporting personalized lifestyle changes.

3. Future Work

We plan to test our systems’ personalization capabilities against a wide array of different user profiles by interacting with LLM-based simulated users\(^7\), and to personalize the recommendations using reinforcement learning \(^8\).

References


\(^2\)Demo video available at: https://youtu.be/slFpI9uBdq4
Institutional Modelling: A Case Study of the Swedish Organ Donation System

Bertilla FABRIS a,1, Michael BELFRAGE a,b and Fabian LORIG a,b
a Department of Computer Science and Media Technology, Malmö University, Sweden
b Internet of Things and People Research Center, Malmö University, Sweden
ORCID ID: Bertilla Fabris https://orcid.org/0009-0000-2551-2692, Michael Belfrage https://orcid.org/0009-0004-1712-5181, Fabian Lorig https://orcid.org/0000-0002-8209-0921

Abstract. Understanding the potential impact of policy changes before implementation is vital, and can be achieved through modelling and simulation. To adequately model stakeholders and regulative constraints, we propose the use of Institutional Grammar to facilitate institutional modelling in Agent-based Social Simulations. We present an early-stage case study exploring the Swedish organ donation system.

Keywords. Agent-Based Social Simulation, Policy Support, Model Formalisation

1. Introduction

Organ donation (OD) systems involve a multitude of stakeholders as well as logistical and medical constraints, disciplined by legal regulations. Hence, introducing new or adapting existing donation policies is challenging as the health of patients could be jeopardized. Agent-Based Social Simulation (ABSS) has proven to be a suitable paradigm for analyzing complex systems within healthcare scenarios and to investigate the potential consequences of policy changes [1,2]. To date, model formalisation does not follow a fixed methodology; its execution relies on the expertise of modellers [3]. Additionally, when developing ABSS aimed at policy support, a focus on scenarios and transparency is required [4]. Institutional modelling is the practice of developing agent-based models leveraging the social aspects of the target system [5]. In this work, we demonstrate how an ABSS model of the Swedish OD system can be formalised through institutional modelling and Institutional Grammar (IG) 2.0 [6].

Institutional modelling is based on institutions, here divided in three types: rules, norms and shared strategies. Institutions can be individuated using IG [7], defining institutional statements from sources such as interviews, field observations and law documents. Institutions are agglomerates of institutional statements, which can be formed of six components: Attribute (actor), Object (target entity), Deontic (enforcement level), Aim (activity), Context (context), and Or else (consequence of violation). The varying level of definition of each statement (i.e. which components are present within it) de-
terminates which type the institution belongs to. The entirety of the target system can be described via institutions [5].

2. Modelling the Organ Donation System of Sweden

Our project “Facilitators and barriers to the use of agent-based social simulation in organ donation” explores the potential of ABSS to support policy making in the Swedish national OD system. The target system presents multiple scenarios for implementing ABSS, such as the recent change to Donation after Circulatory Death, the withdrawal of family veto, new rules concerning organ preservation treatment and a logistical optimization of organ procurement zones. In the system analysis phase, domain knowledge of the infrastructure was sought. This process led to the discovery of relevant policy documents, namely the Swedish Transplantation law 1995:831 and proposition 2021/22:128. We analysed these documents according to IG 2.0 principles. Headings and sections pertaining activities outside the scope of the model (e.g. living donors) were removed, the components were individuated, text split into atomic institutional statements from horizontally nested statements [7]. An additional step necessitated by the nature of the document (Swedish text) was that of translating the original text to English by means of DeepL [8]. The text can then be processed with IG Parser [9]. Thanks to the atomization of statements, areas in which the legal text does not provide sufficient details are discovered and targeted questions to stakeholders can be compiled. We speculate that the same is true for interviews with stakeholders, which can generate a discovery loop between top-down and bottom-up sources. Applying the method to stakeholder interviews will allow agents to deliberate on the top-down regulatory structure of the system, providing a key element of ABSS which is agent autonomy [10].

The method structures the process of model formalisation, which has been methodologically deficient, and increases its traceability, while funneling stakeholder participation in the modelling process. Validation practices might be incorporated within the pipeline in future work. The process has similarities with the one proposed in MAIA [11], but grounds itself in IG 2.0, rather than IAD. Furthermore, the predetermined scope of the project guides modellers to focus on relevant aspects of the law and disregard information not relevant to the model. This approach allows the modellers to investigate the system in a structured manner which is replicable and increases procedural transparency. We propose that the approach can be generalised to other cases once remaining questions are answered: does the method produce blind-spots or error patterns in the resulting simulations? Is there an imbalance between agent behaviour generated via top-down and bottom-up approaches? The increased transparency of the proposed method reinforces trust in the simulation, which is key to incorporating ABSS into data-driven policy making mechanisms [12].

Acknowledgements

This work was supported by a grant from the Knut and Alice Wallenberg Foundation to SciLifeLab for research in Data-driven Life Science, DDLS (KAW 2020.0239) and by the Wallenberg AI, Autonomous Systems and Software Program – Humanity and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation.
References

Keep Gesturing: A Game for Pragmatic Communication

János Adrián GULYÁS a,1, Miklós Máté BADÓ a,1, Kristian FENECH a,2 and András LŐRINCZ a

a Department of Artificial Intelligence, Eötvös Loránd University, Hungary

Abstract. We present an innovative augmented reality game which aims to demonstrate a new dimension of interaction between humans and large language models through non-verbal gesture-based communication. Players collaborate with an LLM-controlled avatar to identify and correct discrepancies in an augmented reality environment, relying solely on non-verbal cues. The game aims to explore the emergence and evolution of pragmatic communication between humans and AI agents.

Keywords. Gesture-Based Communication, Human-AI Interaction, Large Language Model, HoloLens 2, Pragmatic Communication

1. Introduction

Human-machine interaction may occur across many types of communication channels [1], a common form for this channel to take is of natural language. However, natural language in the form of speech and text may not be suitable given restrictions set by the environment such as in noisy and dynamic environments where these modalities may lead to delays and possible misunderstanding. In such situations non-verbal communicative gestures present themselves as a robust alternative communication channel to enable rapid communication. Importantly the meaning behind individual gestures can change depending on the nature and context surrounding the interaction [2].

Our proposed game2 introduces a novel approach to human-machine interaction through gesture-based communication within a game developed for HoloLens 2. While recent work has explored the idea of collaboration between humans and large language models (LLMs), the focus has been on the text domain for example [3,4]. This game poses various collaborative challenges to a human player and a companion driven by an LLM, that requires them to communicate using exclusively hand gestures. The challenges aim to aid the development of a gesture-based pragmatic language between the two participants and will explore how this new language emerges and improves during various play-throughs with different players, and the implications of this phenomenon for the future of human-computer interaction.

1Equal contributors
2Corresponding Author: Kristian Fenech, fenech@inf.elte.hu

https://youtu.be/j_bAw8e0lNU?si=wVB23530QID6bMbU
2. Methodology

Although the game is specifically engineered for human-machine interaction, two humans are also able to complete the game in its entirety. The game takes place in augmented reality, with various interactive objects (puzzle elements) placed in different parts of the room. Interacting with these objects changes their properties and configuration. An object has a specific configuration that it must be set to at the end of the level for a point to be scored. However, only one of the two participants is able to interact with the objects in augmented reality, and only the other party has access to the rules that define the correct configurations. The two parties are not allowed to exchange this information verbally and are restricted to the use of gestures only.

The game features gesture-based back-and-forth interaction between human and a LLM. The hand-gesture approach allows for enough complexity so that a dictionary of gestures that represent various concepts might emerge. Various metrics are calculated, with some being presented to the participants as they complete (or fail) a level. This serves the purpose of encouraging the advancement of their emerging language, as better metrics reflect more effective communication. The presented metrics include efficiency and score, while hidden metrics also include the overall speed of completion, as well as the appearance of complex concepts in the language.

In our implementation, LLMs are instrumental by stepping into the role of one of the players, engaging in the game’s two-way gesture-based communication. The model currently in use for this is OpenAI’s GPT-4 [5]. After receiving an encoded series of hand motions from the players, the LLM interprets these inputs and replies in the same manner. This creates an environment where both parties can use each other’s gestures and ideas to come to an understanding.

An animated character in augmented reality represents the LLM within the game’s virtual environment, serving as its avatar. Gestures received from the LLM are visually depicted through this avatar’s animation. By controlling its avatar, the LLM can intuitively reply to the human player, guiding them to identify and rectify the errors within the virtual room.

We aim to demonstrating that LLMs can understand and contribute to the creation of a new, personalized language, created in collaboration with a human. We believe this will pave the way for more unique and adaptive methods of interacting with artificial intelligence.

3. Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 952026. Additional support provided by the the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory.

References


What Happens to a Robotic Startup While It Evolves

Alexandra HETTICH a,1

aDepartment of Digitalization, Copenhagen Business School

Abstract. Digital transformation (DT) alongside the increasing use of robots in agriculture prompts at behaviors of startups developing them. This poster presents a stable robotic startup AGROS and draws upon the theory of dynamic capabilities (DC) together with perspectives on strategy from IS and management. It is noticed that collaboration with startup research partners cannot be neglected if we seek to understand startup trajectory over time. Several outcomes are possible depending whether it collaborates explicitly, relies unconsciously on collaboration or starts showing the intentions to cognitively collaborate. The contributions to strategy research and IS literature on entrepreneurship and DT are highlighted.

Keywords. Strategy, Startup, Robotics, Dynamic Capabilities, Entrepreneurship

1. Introduction

Despite the differences between types of farms they all face the challenge to produce more and better, with less, while ensuring productivity and improving biodiversity. Researchers still expect more agricultural innovation to stem from robotics, image recognition and smart farming technologies [1,2,3,4]. Yet, robots can be rented or sold, need software updates and adjustments for the use of new components and, finally, offer versatile functionalities for research on crops. Hence, multi-stakeholder approach becomes a necessary extension of the classical company-centered view to account for this complexity rooted in heterogeneous actor types engaged in robotic co-creation (cf. [5]). The current situation on the robotic market resembles a gold rush. “Dozens of autonomous tractors and tool carriers – concepts, prototypes and full production models have appeared in the past decade. Most of these come from start-up companies and research institutes” [6]. Cases are understudied if a robotic network of multi-stakeholder innovation involves a startup. Ref. [7] note that digital innovation refers to “the use of digital technology during the process of innovating…breaking up … vertical industry silos and creating business … where different players come together and innovate by combining and re-combining their digital technology components and knowledge.” Digital entrepreneurship means “the exchange, transfer and acquisition of knowledge through the use of technology to initiate new ways of doing business” [8]. Industry 4.0,
AI race among firms and digital artifacts places tech ventures into a high-paced and ambiguous environment [9,10,11]. In the previous IS research, the scholars’ focus often was on entrepreneur’s individual traits or creation of new ideas by a startup alone [12]. Yet, as startups usually depend on resources of many stakeholders, I choose to focus on forging research collaborations (as one of its DC) and investigate how they are built and maintained over time. **RQ: What are the DC enablers for a collaborating company at a startup phase to guarantee its survival over time?**

2. **Methodology**

Ref. [13] specified four steps for a research design, starting from selecting the case, defining the additional evidence necessary to strengthen the study and jointly analyzing the data from the case and additional evidence. The primary data was collected through interviews and additional evidence was found through press releases, also in archives of the company, news on Google and farmer online resources. Due to the exploratory nature of this work the initial step was to conduct a pilot case-study (ibid) based on several interviews with executives of a robotic startup and stakeholders, which took place between May 2023 and March 2024, were transcribed and coded.

3. **Case and preliminary analysis; from a newborn to semi-mature startup**

AGROS gradually loses characteristics of a newly founded startup and approaches the stage of semi-mature business. Its CROPROBOT platform is based on modularity of the 3 main entities of guidance and control system, tractor frame and implements connection. Today around 50 robots were sold. The bright outlook for AGROS was aptly put by its Board Chairman after the new investment deal in 2020: “**AGROS is in the process of a major transition from being an engineering house to … an internationally leading company within the sustainable plant breeding of the future. With the new round of investment, … [it] gets the opportunity to execute our international commercialization strategy.**”

4. **Conclusions**

I sought to demonstrate how the business of robotic startups depends on DC of forging strong collaborations with research-oriented institutions. Another goal was to shed light on different phases of startup development until it reaches a maturity stage. The case study enables depicting of personal values of management which shape managerial cognition in certain ways [14]. Focusing on research collaborations in robotics, I aim at stressing the role of interactions among actors in the process of digitally inspired innovation. In one of the workshops covering the umbrella of collaborators it was concluded that: “**[w]e introduced the added value of data-exchange between scanning devices – service providers – acting providers and the opportunities of a digital ecosystem between them.**” Thus, the next steps include conducting more interviews with the startup management and its research collaborators to find out the exact attitude changes to research projects and ecosystem perception in the 2nd relative to the 1st phase.
References


The Impact of Mixed-Initiative on Collaboration in Hybrid AI

Leila METHNANI a,1, Adam DAHLGREN LINDSTRÖM a,2, and Virginia DIGNUM a

Umeå University

ORCiD ID: Leila Methnani https://orcid.org/0000-0002-9808-2037, Adam Dahlgren Lindström https://orcid.org/0000-0002-1112-2981, Virginia Dignum https://orcid.org/0000-0001-7409-5813

Abstract. This paper explores the integration of mixed-initiative systems in human-AI teams to improve coordination and communication in Search and Rescue (SAR) scenarios, leveraging dynamic control sharing to enhance operational effectiveness.

Keywords. Human-AI interaction, mixed-initiative systems, search and rescue, team coordination

1. Introduction

The integration of Artificial Intelligence (AI) into collaborative environments has shown promising results in enhancing team effectiveness, particularly in high-stakes environments such as Search and Rescue (SAR) operations. Effective collaboration requires robust communication and dynamic coordination, which are critical when AI systems are involved in decision-making processes. Explainability in AI, especially for mixed-initiative (MI) robotic control, is important for calibrating trust. It not only ensures that AI outcomes can be trusted but also highlights when they cannot be, e.g. signalling when control should be transferred [1]. Moreover, research has demonstrated that mixed-initiative systems, where control is dynamically shared between humans and AI, can significantly improve the adaptability and efficiency of teams [2]. This study builds on the premise that effective human-AI teaming can dramatically enhance operational outcomes in SAR scenarios.

2. Research Approach

To systematically explore the impact of mixed-initiative systems on human-AI teaming, we developed a simulation environment that mirrors the complexities of real-world SAR operations. This environment serves as a testbed for experimenting with various human-AI interaction paradigms.

1Corresponding Author: Leila Methnani, leila.methnani@umu.se
2Corresponding Author: Adam Dahlgren Lindström, dali@cs.umu.se
Our implementation involved designing a series of game-like scenarios that not only simulate the unpredictability and urgency typical of SAR missions but also integrate both autonomous and semi-autonomous robots with human teams. This setup enables continuous interaction and decision-making under pressure, where AI systems are equipped with capabilities to propose actions, request information, and prioritize tasks dynamically. The game was designed as follows: One human player interacts through the command line with an autonomous (reactive planning) robot in a 6 x 6 grid world called the Shyre. The native inhabitants of the Shyre are called Bobbits and are in need of saving from a fire that is spreading across the environment. Trees also populate this grid world; they are obstacles for the robot, but also promote faster spread of the fire in-game. The human and robot have a joint mission: to save the Bobbit from the fire in the Shyre. Each team member can take initiative in the task of navigating the grid world and reaching the Bobbit in time, using two implementations of the MI control: 1) a single switch for toggling control between human-operated and AI-operated modes, and 2) a mixed control mechanism where both AI and human actions are integrated and may simultaneously compete for control.

Our planned user studies will develop several versions of the Bobbit game to explore information dynamics: 1) no proactive information delivery, 2) random proactive information delivery, and 3) context-aware anticipatory information delivery, while allowing information request by the user. Using the same game platform and scenario, we also intend to address what is presented and how, which are identified as important considerations for MI robotic control [1]. Further, we will assess the impact of these information delivery modes on user perception and performance, culminating in a comprehensive model that integrates all aspects of information interaction. This approach leverages cognitive theories to inform system design and improve human-AI collaboration by focusing on high-level decision-making processes in the presence of varying information flows.

We evaluate domain modeling and initiative sharing dynamics using two methods: behavior trees for small-scale clarity and reinforcement learning from human feedback (RLHF) for larger, complex contexts. The game’s simplicity allows initial encoding using behavior trees. However, for scalable, dynamic contexts, we leverage RLHF, effective in fine-tuning LLMs and aligning with Gricean Maxims to foster social, conversational behaviors in small groups. During RLHF, participants guide the learning process by voting on the AI’s anticipatory information pushes, refining a reward model to enhance the AI’s predictive capabilities.

We assess mixed-initiative system performance based on problem-solving efficiency and participant satisfaction, measured through game scores, resource utilization, and cognitive load, with methods including sensory data and interviews. Focus groups with robotic SAR and RLHF experts further refine our approach, aiming for robust, context-aware AI behaviors.

3. Conclusions

The preliminary findings suggest significant potential for mixed-initiative systems to enhance human-AI teaming in complex, dynamic environments like SAR operations, paving the way for future research and implementation.
References


A Hybrid Intelligent Change Management Approach to Generative AI Adoption

Jacob SHERSONa,1
aCenter for Hybrid Intelligence, Dep. of Management, Aarhus University
ORCiD ID: Jacob Sherson https://orcid.org/0000-0001-6048-587X

Abstract. The emergence of no-code AI platforms presents a revolutionary opportunity for democratizing AI, making it accessible for non-technical users to engage in AI-driven innovation [1]. This democratization is particularly evident in the realm of Generative AI (GenAI), which holds substantial promise for creative and problem-solving applications [2]. Despite the wealth of frameworks addressing IT development and adoption [3, 4], the integration of generative AI's disruptive capabilities into existing paradigms remains underexplored, often relegated to the insights of private consultancy reports [5]. This paper aims to address this gap in academic literature within the interdisciplinary concept of Hybrid Intelligence (HI), synthesizing insights from human-centered AI interface development [8], organizational upskilling [9], and the significance of a psychologically safe environment for collaborative innovation [10,11]. Here, we propose a comprehensive framework for integrating generative AI within organizational change management. This framework, highlighting common challenges and opportunities for innovation in GenAI adoption, is informed by product development workshops and training sessions with hundreds of industry leaders in Northern Europe over the past year.

Keywords. Generative AI, Hybrid Intelligence, Human-Centered AI, change management, organizational learning

1. Challenges in Generative AI Adoption

Organizations face numerous obstacles in leveraging Generative AI, including:

- **Inadequate Policy**: Many organizations lack the necessary IT security policies and funding mechanisms for exploratory AI tool adoption, impeding innovation.

- **Misaligned Expectations**: An IT-centric approach often leads to initiatives that fail to garner widespread support, stifling holistic organizational innovation.

- **Training Gaps**: Simplistic training programs fail to equip employees with the skills needed for effective GenAI utilization, leading to inefficient use of AI outputs.

- **Leadership vision**: Lack of a concrete leadership response to popular tech-narratives of imminent job loss

1 Corresponding Author: Jacob Sherson, sherson@mgmt.au.dk.
2. Framework for HI-powered Generative AI Adoption

Overarching considerations:
1. **Human-controlled assistants**: in the GenAI future, all employees will become high-level task-operators of complex Virtual Assistants (VA)
2. **Empowering Employees as Innovators**: Due to the unprecedented accessibility of GenAI, any employee can become a user and an innovator with GenAI, Fig1a)
3. **Culture today, disruption tomorrow**: Rather than identifying your organizational killer VA application of GenAI, start by building a culture of upskilling and micro-innovations, Fig1b)
4. **Complementary role of GenAI**: GenAI can assist in any task but take over very few

Successful GenAI organizations rely on deliberate and simultaneous emphasis on three key factors:

**a) Product innovation**: We advocate for a graduated approach to GenAI integration, starting with simple prompt-based tasks and progressing to complex data-driven applications, Fig1b). Each innovation type should progress through three stages of customization: i) Adoption of an external prompt or service without modification, ii) Adaptation of the innovation to the organizational context via text or functionality edits and iii) Training through iterative application development involving several learning loops of generating output, evaluating/rating them and learning from failure cases. The latter can only be achieved within a Human-Machine-Learning-Loop (HMLL) which involves an efficient decomposition of tasks into machine and human actions, Fig1c). For this, we recommend applying the notion of prediction (manipulation of large amounts of well-structured data) and judgment (manipulation of large amounts of tacit or contextual knowledge), respectively [12].

**b) Cultural innovation**: A few highlights of the identified steps towards creating a comprehensive and inclusive GenAI innovation culture.

**HI prompt engineering** understanding both strengths and weaknesses of LLMs and changing from an emphasis on rapid LLM-output production and subsequent iterative improvements to a more deliberate task description and calibration with the AI-partner.
Sharing and celebrating nano and micro-innovations organization-wide to create an inclusive culture of innovation in which any employee can contribute. This includes an organization-wide prompt library both as concrete process improvement as well as a part of communicating leadership commitment to upskilling and organizational learning. Including explicit emphasis on prompt curation: a continuous and comprehensive effort to invoke a cultural change in mindset.

ChatGPT@Home: Organize an after-work social lecture on utilizing GPT technologies at home, such as fixing a bike with chatGPT image recognition assistance [13].

c) Strategic innovation highlights:
Human-centered vision: Formulate an easily communicable vision of how the transformation will benefit not just the company but also the employees, e.g. shifting tasks from prediction to judgment or that GenAI is perfectly suited to bolster existing strategic competences.

An innovation friendly data security strategy. In particular introducing a middle layer, where a risk assessment of the potential of business value loss is low, is crucial for climbing efficiently up the GenAI product innovation ladder. Fig1d)

GenAI innovation fund allowing employees easy access to experimentation tools across the GenAI modality spectrum.

Acknowledgements: Research was supported by Copenhagen Fintech as part of the “National Position of Strength programme for Finance & Fintech” funded by the Danish Ministry of Higher Education and Science” and Aarhus University Research Foundation.

References

Operational Criteria of Hybrid Intelligence for Generative AI Virtual Assistants

Jacob SHERSON1, Janet RAFNER and Safinaz Büyükgüzel

Center for Hybrid Intelligence, Dep. of Management, Aarhus University

ORCiD ID: Jacob Sherson https://orcid.org/0000-0001-6048-587X, Janet Rafner https://orcid.org/0000-0001-9264-3334, Safinaz Büyükgüzel https://orcid.org/0000-0002-4692-5143

Abstract. The concept of Hybrid Intelligence (HI) is frequently used interchangeably with Human-Centered AI (HCAI) and more broadly as human-in-the-loop. Dellerman et al. [1] outlined three differentiation criteria, emphasizing in particular the need for an evolving continuum of human-AI learning, a concept that has proven challenging to operationalize effectively. Recent efforts aim to expand the definition of HI beyond the domain of human-computer interaction to include application-oriented insights from management science [2]. This broader perspective integrates vital components such as facilitating end-user co-creation through narrative frameworks that foster psychological safety by addressing fears of job displacement [3,4], mitigating risks of deskilling during system deployment and scaling [5], and supporting business process innovation [2]. Additionally, in contrast to HCAI, the name hybrid intelligence conveys the possibly symmetric human-machine relationship and thereby preserves some of the disruptive potential of automated AI rather than relying on purely augmentation of human tasks and intentions [3]. Explicitly, the HI interaction should not only augment the existing, predefined task but also support aspects such as (business) process and business model re-engineering. Despite these considerations, a thorough discussion on which of the many established HCAI concepts and design guidelines form crucial components in achieving the aims of HI has so far been absent in literature. In particular, as it is becoming more and more likely that most knowledge workers will within a short timeframe become operators of complex virtual assistants tapping into LLMs and natural language interfaces, it becomes urgent to ensure that the human-AI interface and associated narrative is constructed to support HI principles and objectives. To initiate this discussion, we formulate explicitly updated HI design criteria in particular for generative AI virtual assistant design and discuss relevant HCAI concept.

Keywords. Keyword, keyword

1. HI Design Criteria

1 Interaction objective: Design an interaction that empowers users to reimagine and innovate workflows as well as entire task purposes and intended value generation streams while preventing long term effects of employee deskilling due to excessive technological reliance.

1 Corresponding Author: Jacob Sherson, sherson@mgmt.au.dk.
2: Interaction narrative: Incorporate these human-centered objectives explicitly into the overall application narrative and visual design to facilitate psychological safety ensuring end-user engagement into the co-creation of the design process as well as adaptation and continuous innovation after tool deployment. Note, a use case is only truly hybrid intelligent if it can be used as a case for a positive Future of Work for humanity.

3: Interaction design: Design for a continually evolving mutual human-AI learning while realizing high degrees of automation and sense of human control.

The aim is to evolve HI design into a prescriptive set of design principles that application developers can strive to achieve as well as allow for concrete discussions of phenomena that are absent in order to differentiate clearly from the more generically applied concepts such as human-in-the-loop.

Some design concepts and challenges that need to be deliberately and contextually designed for in a hybrid intelligence VAs include:

Narrative explicitation and Interactive visualization to incorporate human concerns into an HCAI loop [6]. This includes taking concerns such as fear of job loss explicitly into the naming and design of virtual assistants. For instance, although presented with a friendly graphic the name Einstein GPT of the Salesforce virtual assistant may not instill a sense of equal partnership: when a query fails, whose fault is it? Einstens or the potentially technologically inexperienced and fearful end user?

Seamless vs seamful design: whereas seamless interaction can enhance a sense of user-flow, it can also hide algorithmic complexities within a black box. Conversely, within seamless design one can anticipate potential breakdowns and exploitatively craft custom solutions and procedural shortcuts connecting and traversing seams [7]. An application of this would be transforming the “undo-button event” in which the user cancels an attempt of automation from a frustrating interruption into a positive, collective learning experience.

Navigate the transition from producer to evaluator to avoid productivity loss by i.a. (i) continuous feedback, (ii) system personalization, (iii) ecological interface design, (iv) main task stabilization and timing, and (v) clear task allocation [8]. A critical HI design challenge involves facilitating users to navigate and experiment with task flows and loops that were unfeasible before Generative AI due to the extensive manual effort required. For example, shifting the focus from detailed writing processes to more strategic planning and outlining activities.

Convergent vs divergent support: Designers explicitly need to consider on a case-by-case basis if the assistant should support divergent or convergent thinking by presenting options or suggested solutions. This could involve showing several options for the next step vs showing/completing several steps along the workflow. Here, attention should in particular be given to the pleasing vs provoking tradeoff either fulfilling expectations or instilling new ideas [9]. The user provocation can be particularly relevant for overcoming confirmation bias and, in the HI context, in supporting task objective re-engineering.
Level of foresight: The phenomenon in human-AI interaction where the AI anticipates the user's desires based on ambiguous input and performs several automated steps in one action is referred to as proactive automation or predictive interaction [10]. This involves the AI system analyzing patterns in user behavior, preferences, and context to predict and execute tasks without explicit instructions at every step, aiming to streamline the user experience by reducing the need for direct commands or inputs.

Acknowledgements: The research project was supported by Copenhagen Fintech as part of the “National Position of Strength programme for Finans & Fintech” funded by the Danish Ministry of Higher Education and Science” and Aarhus University Research Foundation.

References
Facilitating Human Feedback for GenAI Prompt Optimization

Jacob SHERSONa,1 and Florent VINCONb

a Center for Hybrid Intelligence, Dep. of Management, Aarhus University
b Université Paris-Cité and Gustave Eiffel, LAPEA, F-92100
ORCiD ID: Jacob Sherson https://orcid.org/0000-0001-6048-587X, Florent Vinchon https://orcid.org/0000-0003-3443-3683

Abstract. This study investigates the optimization of Generative AI (GenAI) systems through human feedback, focusing on how varying feedback mechanisms influence the quality of GenAI outputs. We devised a Human-AI training loop where 32 students, divided into two groups, evaluated AI-generated responses based on a single prompt. One group assessed a single output, while the other compared two outputs. Preliminary results from this small-scale experiment suggest that comparative feedback might encourage more nuanced evaluations, highlighting the potential for improved human-AI collaboration in prompt optimization. Future research with larger samples is recommended to validate these findings and further explore effective feedback strategies for GenAI systems.

Keywords. Generative AI, prompt engineering, human-machine learning, education

1. Introduction

The incorporation of Generative AI (GenAI) technologies within organizational and research environments emphasizes the urgent need to maximize the performance of these systems to reflect intricate human knowledge and contextual subtleties. Although there is a widespread agreement on the usefulness of GenAI models, optimizing prompts remains a sophisticated challenge that necessitates a thorough examination of the mechanisms that enable productive human-AI interaction.

It is generally essential for a human to be at the center of the improvement loop in order to evaluate the quality of the output produced by a generative AI system like ChatGPT. A human is better equipped to comprehend the context and the circumstances, and can exhibit a refined expertise in the concepts presented, while also adopting a more distinctive and often more pertinent style.

Thus, we recommend the establishment of a human-AI training loop, which begins with a single prompt that is refined and improved over time. In this process, the human plays a pivotal role by evaluating the output and determining its fine-grained properties. By drawing upon their expertise and sensitivity, humans can identify what aspects of the output are more useful and effective. The integration of critical feedback and knowledge of AI-type virtual assistants by experts is crucial to enhancing the quality of Gen AI outputs. This, in turn, leads to a virtuous cycle of continuous skill improvement between

1 Corresponding Author: Jacob Sherson, sherson@mgmt.au.dk.
AI and humans, ultimately resulting in better acceptance and feedback from users and companies.

However, it is often difficult for domain experts to explicitly specify their knowledge within their field of expertise. Indeed, expressing a nuanced and contrasting opinion on anything requires comparison. As it stands, there are no "golden standards" for evaluating these prompts, which is why we wanted to propose several readings for evaluating the effectiveness of a prompt, inspired by studies on absolute and relative performance [4, 5, 6]. In this type of study, it is generally emphasized that there are several examples of a task to be performed (for us, the identical prompt) in the evaluation of situations or texts [2, 3].

Tools like ChatGPT are able to facilitate this seamlessly, offering us several solutions and letting us choose which one we prefer to suit our style. However, this strategy does not provide viable answers immediately, which could be improved by the virtuous loop discussed earlier.

Therefore, we have conducted a pilot study with two conditions, both with the same general objective: to read one or more outputs from a generative AI and to evaluate them in two groups. For the first group, the aim was to carry out a set of systematic evaluations of the results of a prompt: reading the prompt, evaluating it on a scale ranging from 1 (very bad) to 5 (very good), then describing their reflections, clearly stating the positive and negative elements, and distinguishing avenues for improvement from the commentary. In the second group, the aim is to study several outputs that are still based on the same Gen-AI and the same prompt. This is in line with Shah's [7] remarks on how to make prompt and outcome studies more scientific. In cases where concrete contextual feedback is required, users will be more able and motivated to provide it if they are presented with two outcomes from the same prompts and asked to rate and comment on both on an absolute scale, as opposed to the rule of displaying only one outcome to rating at a time?

We conducted an experiment with 32 Danish high-school students divided into two groups to examine the impact of varying prompts on their critical response skills. The prompt was designed to be relevant to problems that students may encounter in real life, and aimed to assess the feasibility of incorporating AI into their daily school activities. Based on the overall school-task, the prompt produced a detailed task breakdown and for each made suggestions for the potential level and method of incorporating AI into the subtask solution giving a human-AI task hybridization level ranging from 1 (human performed the majority of the task) to 5 (AI performed the majority of the task). This prompt was used multiple times in different ChatGPT4 conversations to ensure the presence of two distinct AI individuals. The first group (NGr1 = 19) received only one response, while the second group (NGr2 = 13) received two, one identical to the first group and the other different. Both groups were asked to describe the outcome, rate it on a scale of 1 to 5 (1 being very bad and 5 being excellent) and provide justification for their scores.

Although our sample size was small, we found it interesting to examine the descriptive differences between the two groups. Most notably there was an inverse difference in the number of words used, with Group 1 using fewer words (mean of 21) than Group 2 (mean of 26), which might indeed suggest that Group 2 engaged more deliberately in the evaluation process. Because additional research is necessary to establish a definitive conclusion, we believe that these initial descriptive findings are worth exploring. It is possible that the number of words used is an indicator of the quality of a student's justification or argument [1].
Despite the fact that our research did not yield statistically significant outcomes as a result of the limited number of participants, it is noteworthy that, at least in terms of descriptive statistics, disparities do exist and appear to support our hypothesis. Further investigation could be conducted by engaging larger sample sizes, examining participants’ critical abilities, broadening the demographic scope, and considering additional variables.

**Acknowledgements:** The research project was supported by Copenhagen Fintech as part of the “National Position of Strength programme for Finans & Fintech” funded by the Danish Ministry of Higher Education and Science” and Aarhus University Research Foundation.

**References**


Value-Sensitive Disagreement Analysis for Online Deliberation

Michiel VAN DER MEER\textsuperscript{a,1}, Piek VOSSEN\textsuperscript{b}, Catholijn M. JONKER\textsuperscript{c}, and Pradeep K. MURUKANNAIAH\textsuperscript{c}

\textsuperscript{a}LIACS, Leiden University
\textsuperscript{b}CLTL, Vrije Universiteit Amsterdam
\textsuperscript{c}II, TU Delft

Abstract. Disagreements are common in online societal deliberation and may be crucial for effective collaboration, for instance in helping users understand opposing viewpoints. Although there exist automated methods for recognizing disagreement, a deeper understanding of factors that influence disagreement is currently missing. We investigate a hypothesis that differences in personal values influence disagreement in online discussions. Using Large Language Models (LLMs) for estimating both profiles of personal values and disagreement, we conduct a large-scale experiment involving 11.4M user comments. We find that the dissimilarity of value profiles correlates with disagreement only in specific cases, but that incorporating self-reported value profiles changes these results to be more undecided.

Keywords. perspectives, values, natural language processing, hybrid intelligence

1. Introduction

A large number of users participate in online deliberations on societal issues such as climate change [1] and vaccination hesitancy [2]. Disagreement is important to deliberation [3,4], since it helps users understand the opposing viewpoints [5,6]. Measuring disagreements may therefore be useful for measuring deliberation quality [7]. However, recognizing the level of disagreement in a discussion is hard because a deeper understanding of factors that influence disagreement is currently missing [8].

To a degree, automated methods can be used to analyze disagreement in opinions expressed through text. Contemporary approaches are usually based on LLMs, and encompass a variety of tasks like Sentiment Analysis [9], Stance Detection [10], and Argument Mining [11]. However, these methods are shown to disregard diverse opinions [12,13,14], and only paint a partial image of the opinions at play. Analyses of deliberations require a more accurate description of the diverse opinions to be effective [15,16].

We propose to instead use a proxy for disagreement by focusing on value conflicts, a potential root cause for disagreement [17]. Preferences among values are used for representing the motivations underlying opinions and actions. In this abstract, we briefly describe how we tested the hypothesis that when users with conflicting values engage in a discussion, diverging views come up [18].

\textsuperscript{1}Corresponding author: Michiel van der Meer, m.t.van.der.meer@liacs.leidenuniv.nl.
2. Value-Sensitive Disagreement Analysis

To evaluate our hypothesis, we estimate individual value profiles based on user comments on Reddit, a social media platform. A value profile captures the relative importance a user ascribes to values from the well-known Schwartz theory of basic values [19]. We compare the similarities among profiles to cases of (dis-)agreement among users. This allows us to investigate the association between value conflict (low similarity) and disagreement. Figure 1 shows a detailed overview of our approach. We train a classifier to perform Value Extraction on individual user comments. We aggregate predictions for a single user and construct the value profile by scoring the relative frequency of value mentions. We test if there is a significant difference among the profiles of user pairs that have been shown to agree with each other versus those that disagree with each other.

We gather 11.4M comments from 19K users on Reddit to construct value profiles. We perform up to 200 tests with different settings to investigate our hypothesis. We further experiment with replacing estimated value profiles with self-reported ones. To do so, we collect 572 judgments from 26 annotators in combination with self-reported value profiles. We mostly observe negative evidence of a correlation between profile similarity and disagreeing users when using VPE. This corroborates the difficulty of solely automated discussion analysis with LLMs. When including self-reports, the results change for a majority of the cases, alluding to differences in how profiles are obtained. The specific cases in which value conflicts correlate with disagreements are those where values are likely to be relevant and diverse. Therefore, while value conflicts may not be directly related to disagreement, they signal the diversity of the underlying motivations of participants. How context influences the meaning and relevance of values is usually unknown, though Hybrid Intelligence (HI) approaches can be used to inform decision-making [20]. This opens up human-machine collaboration opportunities for a more constructive conversation [21,22,23].


ChatGPT May Foster Human Gist Memory While Offloading Less Sufficient Information

Gleb VZORIN a,1, Vasilissa PETROVA a and Anna SEDYKH a
Lomonosov Moscow State University
ORCiD ID: Gleb Vzorin https://orcid.org/0000-0003-2034-8007, Vasilissa Petrova https://orcid.org/0000-0001-7021-5942, Anna Sedykh https://orcid.org/0000-0000-9950-6906

Abstract. Recent research challenges the notion that digital offloading impairs memory, suggesting instead that at least gist memory may stay intact. Drawing from Walter Kintsch’s tripartite model of text comprehension, we hypothesize that offloading spares cognitive resources, thus not merely sparing gist memory, rather improving it. Our methodology involves two experiments testing memory retention across different levels of text comprehension, under conditions of anticipated text accessibility, including the use of ChatGPT. Preliminary results from these experiments indicate comprehension level effect, with higher-order cognitive processes benefiting more from offloading. Although pilot data has no statistical significance, they hint at a potential advantage of digital offloading for gist memory, challenging traditional perceptions of memory decline due to technological reliance. The ongoing research aims to further elucidate the impact of digital offloading on memory processes, potentially shaping future hybrid cognitive architectures where human creativity and AI-driven memory retention coalesce.

Keywords. Cognitive offloading, Google-effect, digital externalization, hybrid intelligence.

Experimental evidence suggests that individuals are more likely to forget information that they have saved digitally, whereas information that has been deleted is more likely to be remembered [1]. This tendency to offload information is often perceived as a general decline in memory capacity. Recent experiments, however, reveal inconsistencies in this observed decline. A series of studies [2] demonstrated that the advantage in recalling categorized words, as opposed to non-categorized ones, was either equivalent to or greater in the offloading condition than in a condition where participants did not anticipate having access to their written lists during the recall test. Therefore, it appears that gist memory, which is bolstered by semantic associations, does not undergo significant reduction when offloading.

Here we propose a stronger hypothesis that gist memory may be not merely intact, but rather increased, because the cognitive resource, spared by offloading, is redirected from memorizing towards comprehension. To address this improvement, we need to use

1 Corresponding Author: Gleb Vzorin, g.vzorin@mail.ru
more detailed approach to levels of memory. The proper theoretical framework was proposed by Walter Kintsch [3], who suggested the tripartite model of text comprehension: text surface (TS), textbase (TB), and situation model (SM). The TS refers to the exact wording, the TB covers the semantic content that can be seen as a network of propositions, the SM is a coherent representation of the situation referred to in the text and is constructed by drawing inferences.

To examine our hypothesis, we conducted two experiments, each comprising of text reading and test stages. The instructions in the first stage present three conditions about text accessibility in the following test stage: 1) the text will not be available, 2) the text will be available, and 3) ChatGPT can be used for task resolution. Participants are students well-versed in ChatGPT’s functionalities. In both experiments, during the testing phase, participants were, in fact, denied access to the text, allowing for the assessment of actual memory retention.

The first experiment employed sentence recognition paradigm [4] to access memory separately for three levels. Participants read 12 short texts on different topics under three aforementioned conditions. On the text day, they are presented with separate sentences from the text and asked to decide whether this sentence had appeared in the original text. These sentences could be either verbatim from the text or modified to align with one of the three memory levels being tested.

In the second experiment, we refined the textbase level into two sub-levels to differentiate between micro and macro semantic comprehension. During the initial stage, participants were exposed to three texts. In the testing stage, they were presented with multiple-choice questions designed to directly assess comprehension at one of the four specified levels. This approach allowed us to more precisely evaluate the nuances of semantic understanding facilitated by the offloading process.

Analysis of pilot series (N=27) from the second experiment yielded inconclusive results that, while not statistically significant, tended to support our hypothesis. A two-way repeated measures ANOVA revealed a significant effect for the Comprehension level (p < 0.001), but no significant interaction between Offloading type and Comprehension level (F(6, 26) = 1.67, p = 0.131). Absolute mean values show that in “ChatGPT” performance in macrosemantic and SM levels was highest, while in microsemantic and surface it was poorest. The performance in no-offloading condition was greatest on microsemantic level. To assess the quality of the questions, we administered them to individuals who had not read the texts. Their responses did not surpass the level of random guessing, indicating that the formulation of the questions was neutral and did not prime the answers.

Due to the high variance in the data and the small sample size in the pilot study, we were unable to obtain statistically significant results. However, the general patterns observed in the results align with our hypothesis, suggesting that AI and digital offloading might not necessarily impair human memory. Instead, they may offer benefits for higher-order gist memory of texts. The forthcoming results from the main experimental series are expected to provide significant insights into hybrid cognitive architectures. In these architectures, the role of humans is envisioned to be more creative, focusing on the management of generalized ideas, while AI fosters information retention.
References


AI + Object Making & Creative Content

Kio ZHU and William NICKLEY
The Ohio State University

Abstract. Research in AI inspired instrument making and content creation. This study explores the integration of Artificial Intelligence (AI) in the creation of artistic and musical objects, employing hybrid human-machine processes. AI technologies, including text generation, text-to-picture, and the innovative text-to-3D, are rapidly evolving and becoming more accessible to artists and designers. Our project investigates the potential of these technologies to enhance creative processes in the arts and design.

Keywords. Hybrid human/machine process, object design, instrument, virtual content creation

1. Introduction

The project is an investigation into what it means culturally for AI to make creative contents, which is fundamentally human centered. How would AI interpret abstract concepts? How do we as humans assign meaning to AI generated contents?

The inception of our project was inspired by Dream- field, a text-to-3D model developed by researchers at UC Berkeley and Google Research, including Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Since then, the field has advanced with the introduction of several models, including Stable-Dreamfusion, which is based on Google’s Dreamfusion and incorporates Stable-Diffusion techniques. Our work has led to the creation of physical objects that embody the sonic characteristics of various musical artists, through a complex, collaborative, hybrid human-machine process. This process involves generating descriptive prompts with ChatGPT, creating virtual 3D objects with Stable-Dreamfusion, and transforming these into tactile, interactive objects using virtual reality and 3D modeling technologies.

2. Research

Our methodology involves multiple stages: conducting interviews with musicians, generating textual descriptions of their sound, creating 3D models of these descriptions, manually refining the models, prototyping through 3D printing, and ultimately producing real-scale, interactive musical instruments with sonic, physical, and visual components.

An initial focus has been the production of both physical and virtual representations of sound artists’ work, incorporating feedback from local, underground music scenes to enrich our creative process. Collaborations with local bands such as Catchwords and Abel have been instrumental in this endeavor. Future phases will focus on refining our outputs to fully functional musical instruments suited for live performances, and
exploring the development of new AI-driven 3D modeling techniques that go beyond textual inputs.

Our project also brings into focus cultural implications of AI in creative content generation, questioning how AI interprets abstract concepts and how humans attribute meaning to AI-generated content. This exploration is crucial in understanding the broader impact of AI on creative industries associated with art and design.

3. **Presentation and showcase**

Imagine being drawn by intriguing sounds to a corner where a dazzling display awaits. You see individuals skillfully manipulating beautiful, strange objects, almost dancing in sync with the captivating music they control. Nearby, another object invites you to join this symphony. This isn't just a demonstration—it's an invitation to engage directly with our AI-inspired creations, where technology meets human expression. Come explore how these objects transform sounds and visuals in real-time, inviting you to be part of the artistry.

During the workshop, we aim to demonstrate our streamlined process for creating AI-inspired musical objects and allow participants to engage with this process firsthand. Additionally, we are constructing an interactive installation where attendees can interact with musical object prototypes, triggering AI-generated sounds and visuals. This immersive experience is designed to engage participants directly with the outcomes of our research and explore the intersection of AI and human creativity.

4. **Link to video demo**

https://youtu.be/Hv9k4VJr_mA

5. **Special Thanks**

Grace Gerber as 3D-modeler, a human in our hybrid process. Catchwords (band) and Abel (band) as musician collaborators, humans in our process.

6. **AI models used**

## Author Index

<table>
<thead>
<tr>
<th>Name</th>
<th>Page Numbers</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abramski, K.</td>
<td>3, 451</td>
<td>Fleiner, C. 124</td>
</tr>
<tr>
<td>Adhikari, A.</td>
<td>13</td>
<td>Ganapathy, S. 105</td>
</tr>
<tr>
<td>Akintunde, M.</td>
<td>22, 335</td>
<td>Gezici, G. 64</td>
</tr>
<tr>
<td>Alfieri, C.</td>
<td>36</td>
<td>Ghebremedhin, S. 145</td>
</tr>
<tr>
<td>Armaselu, F.</td>
<td>46</td>
<td>Giannotti, F. 64</td>
</tr>
<tr>
<td>Auda, J.</td>
<td>213</td>
<td>Gruenefeld, U. 213</td>
</tr>
<tr>
<td>Avila, K.E.</td>
<td>199</td>
<td>Gulyás, J.A. 463</td>
</tr>
<tr>
<td>Aydin, H.</td>
<td>55</td>
<td>Gürçan, Ö. 134</td>
</tr>
<tr>
<td>Badó, M.M.</td>
<td>463</td>
<td>Haliburton, L. 145</td>
</tr>
<tr>
<td>Báez Santamaría, S.</td>
<td>315</td>
<td>Hansson, J. 325</td>
</tr>
<tr>
<td>Baez Santamaria, S.</td>
<td>457</td>
<td>Hartikainen, M. 162, 371</td>
</tr>
<tr>
<td>Baier, T.</td>
<td>315</td>
<td>Hettich, A. 466</td>
</tr>
<tr>
<td>Belfrage, M.</td>
<td>460</td>
<td>Heylen, D. 173, 350</td>
</tr>
<tr>
<td>Ben Allouch, S.</td>
<td>84</td>
<td>Ilesanmi, F. 241</td>
</tr>
<tr>
<td>Ben Chaaben, E.</td>
<td>363</td>
<td>Inverardi, P. 36</td>
</tr>
<tr>
<td>Beretta, B.</td>
<td>188</td>
<td>Járvelä, S. 241</td>
</tr>
<tr>
<td>Beretta, I.</td>
<td>73</td>
<td>Javadani Rikhtehgar, D. 173</td>
</tr>
<tr>
<td>Berman, A.</td>
<td>454</td>
<td>Jonker, C.M. 481</td>
</tr>
<tr>
<td>Borglund, E.</td>
<td>325</td>
<td>Kamphorst, B.A. 457</td>
</tr>
<tr>
<td>Brännström, A.</td>
<td>325</td>
<td>Karolus, J. 199</td>
</tr>
<tr>
<td>Büyükgüzél, S.</td>
<td>475</td>
<td>Kiefer-Emmanouilidis, M. 199</td>
</tr>
<tr>
<td>Cappuccio, E.</td>
<td>64, 73</td>
<td>Kox, E.S. 188</td>
</tr>
<tr>
<td>Chang, W.-T.</td>
<td>84</td>
<td>Krämér, N. 264</td>
</tr>
<tr>
<td>Chen, P.-Y.</td>
<td>457</td>
<td>Kramer, S. 84</td>
</tr>
<tr>
<td>Cirstea, C.</td>
<td>22</td>
<td>Krupp, L. 199</td>
</tr>
<tr>
<td>Citraro, S.</td>
<td>451</td>
<td>Küchemann, S. 199</td>
</tr>
<tr>
<td>Crescenzi, P.</td>
<td>64</td>
<td>Kuhn, J. 199</td>
</tr>
<tr>
<td>Dahlgren Lindström, A.</td>
<td>v, 469</td>
<td>Kühn, T. 264</td>
</tr>
<tr>
<td>Dang, B.</td>
<td>241</td>
<td>Lavorati, C. 3</td>
</tr>
<tr>
<td>Dastani, M.</td>
<td>22</td>
<td>Liebers, C. 213</td>
</tr>
<tr>
<td>Dave, B.</td>
<td>105</td>
<td>Lombardi, L. 451</td>
</tr>
<tr>
<td>de Boer, M.H.T.</td>
<td>457</td>
<td>Lorig, F. v, 460</td>
</tr>
<tr>
<td>de Boer, V.</td>
<td>94, 350</td>
<td>Lörincz, A. 463</td>
</tr>
<tr>
<td>De Sanctis, M.</td>
<td>36</td>
<td>Lukowicz, P. 199</td>
</tr>
<tr>
<td>den Hengst, F.</td>
<td>457</td>
<td>Maathuis, H. 379</td>
</tr>
<tr>
<td>Dhondt, S.</td>
<td>13</td>
<td>Mac Namee, B. 402</td>
</tr>
<tr>
<td>Dignum, F.</td>
<td>v, 114</td>
<td>Marchiori Manerba, M. 73</td>
</tr>
<tr>
<td>Dignum, V.</td>
<td>469</td>
<td>Marcu, A. 300</td>
</tr>
<tr>
<td>Donati, D.</td>
<td>36</td>
<td>Mayer, S. 145</td>
</tr>
<tr>
<td>Ellis, A.</td>
<td>105</td>
<td>McKinlay, J. 386</td>
</tr>
<tr>
<td>Erdogan, E.</td>
<td>114</td>
<td>Methnani, L. 469</td>
</tr>
<tr>
<td>Fabris, B.</td>
<td>460</td>
<td>Meyer-Vitali, A. 228</td>
</tr>
<tr>
<td>Fenech, K.</td>
<td>463</td>
<td>Monaldini, A. 394</td>
</tr>
<tr>
<td>Name</td>
<td>Page Numbers</td>
<td>Name</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Moreau, L.</td>
<td>22, 335</td>
<td>Steinert, S.</td>
</tr>
<tr>
<td>Mulder, W.</td>
<td>228</td>
<td>Stella, M.</td>
</tr>
<tr>
<td>Murukanahiah, P.</td>
<td>v</td>
<td>Stork, L.</td>
</tr>
<tr>
<td>Murukanahiah, P.K.</td>
<td>481</td>
<td>Theodorou, A.</td>
</tr>
<tr>
<td>Neerinex, M.A.</td>
<td>300</td>
<td>Tiddi, I.</td>
</tr>
<tr>
<td>Nguyen, A.</td>
<td>241</td>
<td>Tielman, M.L.</td>
</tr>
<tr>
<td>Nickley, W.</td>
<td>488</td>
<td>Tucker, J.</td>
</tr>
<tr>
<td>Nieves, J.C.</td>
<td>325</td>
<td>Vaānānen, K.</td>
</tr>
<tr>
<td>Oey, M.</td>
<td>84</td>
<td>van der Meer, M.</td>
</tr>
<tr>
<td>Parise, A.</td>
<td>402</td>
<td>van der Meulen, R.</td>
</tr>
<tr>
<td>Peng, X.</td>
<td>409</td>
<td>van Duijn, M.</td>
</tr>
<tr>
<td>Petrova, V.</td>
<td>485</td>
<td>van Genabeek, J.A.G.M.</td>
</tr>
<tr>
<td>Pfützenreuter, N.</td>
<td>213</td>
<td>van Leeuwen, L.</td>
</tr>
<tr>
<td>Quadt, T.</td>
<td>418</td>
<td>van Riemsdijk, M.B.</td>
</tr>
<tr>
<td>Rafner, J.</td>
<td>475</td>
<td>Veenman, C.J.</td>
</tr>
<tr>
<td>Ramanath, S.</td>
<td>255</td>
<td>Vennekens, J.</td>
</tr>
<tr>
<td>Ram Rath, M.</td>
<td>264</td>
<td>Verbrugge, R.</td>
</tr>
<tr>
<td>Renooij, S.</td>
<td>291</td>
<td>Verhagen, R.S.</td>
</tr>
<tr>
<td>Ridder, A.</td>
<td>264</td>
<td>Verheij, B.</td>
</tr>
<tr>
<td>Rinzivillo, S.</td>
<td>64, 73</td>
<td>Vinchon, F.</td>
</tr>
<tr>
<td>Rossetti, G.</td>
<td>3, 451</td>
<td>Vossen, P.</td>
</tr>
<tr>
<td>Salehi Fathabadi, A.</td>
<td>22</td>
<td>Vossers, J.</td>
</tr>
<tr>
<td>Salehi, H.</td>
<td>105</td>
<td>Vuorenmaa, E.</td>
</tr>
<tr>
<td>Scharmann, A.</td>
<td>264</td>
<td>Vzorin, G.</td>
</tr>
<tr>
<td>Schlobach, S.</td>
<td>173</td>
<td>Wang, S.</td>
</tr>
<tr>
<td>Schmidt, A.</td>
<td>145</td>
<td>Weerawardhana, S.</td>
</tr>
<tr>
<td>Schneegass, S.</td>
<td>213</td>
<td>Weller, S.</td>
</tr>
<tr>
<td>Sedykh, A.</td>
<td>485</td>
<td>Welsch, R.</td>
</tr>
<tr>
<td>Serafin, P.B.</td>
<td>64</td>
<td>Wolff, J.</td>
</tr>
<tr>
<td>Sherson, J.</td>
<td>472, 475, 478</td>
<td>Yazdanpanah, V.</td>
</tr>
<tr>
<td>Shimizu, C.</td>
<td>105</td>
<td>Yolum, P.</td>
</tr>
<tr>
<td>Smit, Q.</td>
<td>13, 457</td>
<td>Zamprogno, G.</td>
</tr>
<tr>
<td>Sorysz, J.</td>
<td>427</td>
<td>Zhu, K.</td>
</tr>
<tr>
<td>Spurava, G.</td>
<td>162</td>
<td></td>
</tr>
</tbody>
</table>