# Human in Command Machine Learning

**Lars Holmberg**
Computer Science and Media Technology
Malmö University
Lars.Holmberg@mau.se

## Abstract

Work in Artificial Intelligence and Machine Learning (AI/ML) often focuses on reproducing human intelligence, an approach that can raise concern regarding the possibilities of human flourishing in harmony with AI/ML. In my work, as an alternative, I focus on human-in-command machine learning, a setting where an AI/ML artifact is taught by a human teacher. I use a prototype to gather knowledge as a starting point for an analysis concerning how this approach changes the way machine learning is done.

## Introduction

Contemporary machine learning research is currently dominated by a technocentric perspective that primarily solves and addresses technical and functional goals but risks not placing enough focus on serving human values and ethical principles (IEEE 2019, p. 2). As an alternative, moving the agency towards domain experts and end-users is an emerging and promising field. A move in this direction can help democratize the knowledge and thus mitigate some of the risks of the problematic knowledge concentration of this disruptive technology (Couldry and Mejias 2019). An agency move in the direction of non-machine learning experts also offers the possibility to find and explore additional application areas for the technology.

In the Ethics Guidelines for Trustworthy AI (HLEG 2019, p. 18) a governance mechanism giving different levels of oversight is defined. The levels, with increasing human agency, are named as human-in-the-loop (HITL), human-on-the-loop (HOTL) and human-in-command (HIC). Approaches like Active Learning, Interactive Machine Learning and Machine Teaching involves human domain experts in the training and can be arranged on this governance scale (Zhu et al. 2018; Simard et al. 2017).

In my research, I concentrate on HIC as an approach, by selecting this route I start my explorations from a distinctly human perspective to gain insights from an extreme point of the agency shift. Using this approach in a concrete context I intend to build knowledge that can be transferable between domains. For the current phase in my research, I focus on the following research question.

- How does a human-in-command approach change the way machine learning is done?

To shed light on this question I created a early high-fidelity prototype in the form of a smartphone app, targeting the personal knowledge domain commuting. As a domain, commuting is well-known whereas commute patterns are individual and the commuters themselves are experts in their own patterns.



## Research foundation

HIC is an approach where a human teacher gains control over the transfer of subjectively selected knowledge to a model in an AI/ML system. There is, in this setting, no demand for the user to know machine learning, instead, the human needs relevant domain knowledge. To achieve this a domain-specific language and design has to be developed for the domain (Simard et al. 2017).

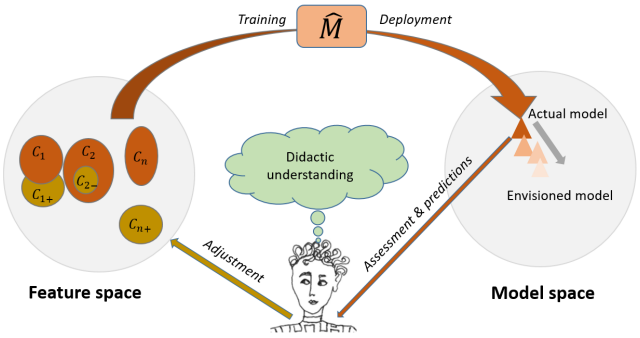


Figure 1: Overview of the HIC system in this work.

In my approach, a human selects sub-spaces in a feature space $C_1, ..., C_n$, spaces used to train a model (figure 1). The model is assessed in order to evaluate if the model's knowledge maps the knowledge the teacher intended to transfer. The teacher tries to move the actual model towards the envisioned in an iterative process where training data is added

or removed ($C_{1+}, C_{2-}, ..., C_{n+}$) and the ML-method $\hat{M}$ retrained until it fulfills the teacher's expectations.

I have used Research through Design with the aim of using a design practice that is "brought to bear on situations chosen for their topical and theoretical potential [...]" (Gaver 2012, p. 937)

## Prototype

My prototype is a fully functional Android app for commuters[1,2]. A commuter can use the prototype to teach the app journeys from their commute patterns, they do this by connecting a time-span, day and location to a journey such as "going to work". By using the commuters context the prototype then can predict upcoming journeys in real-time. A personal ML-model is trained for each user.

## Findings and discussion

In my case study, commuting, the one to one relationship between a human teacher and a machine learning system is in focus. When the prototype is first started by the user no training data exist, instead initial training data is created during teaching sessions.

The need for an initially crude and fast teaching became apparent during our evaluation. Over time the teaching needs are more fine-grained and focused on adjustment of the training data and knowledge assessment. This result is in line with findings in other domains.

It is clear, from my work, that using HIC systems pose a didactic challenge for the user in that the teaching and assessment process has to be humanly understandable. From the perspective of the HIC system designer, the learning process is designed and, depending on the aim of the learning process, experience could be drawn from different educational theories.

Making use of and saving the teacher's domain specific meta-knowledge during teaching is an interesting opportunity that can facilitate system interpretability. This knowledge can for example be expressed as concepts in a relational graph related to areas in the feature space (Simard et al. 2017). An approach following these lines can add a meta-layer to the labeled data. The ability to add and organize this meta-information as concepts is an opportunity that puts demand on the teaching interface.

I also found that constraining model generalization through confining classification boundaries is important for this type of systems. This, in order to design, a for the commuter, logical classification behavior concerning predictions made in areas in the feature space where no training data exists.

The work also suggests that through designing an HIC system for a domain, over time the system can be trusted. To reach this the system has to be either, as in our case, matching predictions to the context in real-time in an interpretable way and/or support exploration regarding the model so the knowledge transferred and behavior can be assessed and understood. We can also see a need for model-agnostic interpretability approaches (Ribeiro, Singh, and Guestrin 2016; Lakkaraju, Bach, and Leskovec 2016).

[1]https://skanependlaren.firebaseapp.com/

[2]https://github.com/k3larra/commuter

## Future directions and plans

Through this work, we, me and my supervisors, gained understanding of some of the challenges and the opportunities that a HIC approach has so we in part answer the main research question. In the next step of our research, we aim at transferring our gained knowledge to a shared knowledge domain, in which we, as in our initial domain can create an artifact that can be evaluated.

One domain that we discuss and we have access to, through a research project with an industrial partner, is district heating. In this domain opportunities to save energy exists by exchanging heating and cooling needs for groups of buildings since it is often easier to use an unnecessary amount of energy instead of risking a too hot or cold building. One possibility is to create a HIC system for each building, using for example janitors as teachers, and create a indoor climate in line with the activities in the building and at the same time save energy for the building group.

We are also discussing other areas and are open for cooperation with other researches and other domains, for example in the area of assistive technology, intelligent personal assistants and personal informatics.

## References


Couldry, N., and Mejias, U. A. 2019. Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media* 20(4):336–349.

Gaver, W. 2012. *What Should We Expect From Research Through Design?*

HLEG. 2019. Ethics Guidelines for Trustworthy AI (European Commission, 2019). Technical report, High-Level Expert Group on Artificial Intelligence.

IEEE. 2019. Ethically aligned design. Technical report.

Lakkaraju, H.; Bach, S. H.; and Leskovec, J. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, 1675–1684.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 1135–1144. New York, New York, USA: ACM Press.

Simard, P. Y.; Amershi, S.; Chickering, D. M.; Pelton, A. E.; Ghorashi, S.; Meek, C.; Ramos, G.; Suh, J.; Verwey, J.; Wang, M.; and Wernsing, J. 2017. Machine Teaching: A New Paradigm for Building Machine Learning Systems. Technical report, Microsoft Research.

Zhu, X.; Singla, A.; Zilles, S.; and Rafferty, A. N. 2018. An Overview of Machine Teaching. *arXiv preprint arXiv:1801.05927*.