Johan Brännmark*

Means Paternalism and the Problem of Indeterminacy

https://doi.org/10.1515/mopp-2021-0032
Published online December 21, 2021

Abstract: Many contemporary defenders of paternalist interventions favor a version of paternalism focused on how people often choose the wrong means given their own ends. This idea is typically justified by empirical results in psychology and behavioral economics. To the extent that paternalist interventions can then target the promotion of goals that can be said to be our own, such interventions are *prima facie* less problematic. One version of this argument starts from the idea that it is meaningful to ascribe to us preferences that we would have if were fully rational, informed and in control over our actions. It is argued here, however, that the very body of empirical results that means paternalists typically rely on also undermines this idea as a robust enough notion. A more modest approach to paternalist interventions, on which such policies are understood as enmeshed with welfare-state policies promoting certain primary goods, is then proposed instead.

Keywords: paternalism, policy-making, rationality, welfare, preferences

1 Introduction

Recent years have seen a resurgence in defenses of government paternalism. The most well-known approach here is the libertarian paternalism developed and defended by Thaler and Sunstein (2003, 2008), which emphasizes the use of *nudges* (interventions that *de facto* influence our choices, but without reducing our freedom of action), whereas other leading writers, such as Conly (2013) and Le Grand and New (2015), are open to, or even prefer, more direct forms of paternalism. One thing uniting these writers is that they tend to take recent findings in behavioral economics and psychology as a starting point (Pickett 2019: 301), and where these empirical results are often taken to show that we are prone to make a variety of mistakes in our everyday choices – mistakes which the right kinds of policy-making can compensate for. As noted by Hanna (2018: 103–04), this line of reasoning often relies on a distinction between *means-related* and *ends-related*

***Corresponding author: Johan Brännmark**, Department of Global Political Studies, Malmö University, Nordenskiöldsgatan 1, 211 19 Malmö, Sweden, E-mail: johan.brannmark@mau.se

paternalism, the idea being that the relevant mistakes occur on the means side of things, and that they are mistakes precisely because they do not line up with our own more general and overarching ends.[1]

In what follows here, it will be argued that while the idea of means paternalism is understandably attractive, it is a position that hinges on being able to articulate a notion of ends against which our everyday choices can be contrasted – but a worry then is that this move will ultimately be undermined by the very literature in psychology and behavioral economics that these contemporary defenders of paternalism tend to appeal to. Given the murky and shifting character of our everyday deliberations, how are we supposed to derive a determinate conception of ends that are supposed to be our real ends or true preferences? One possible response is that this problem is just an *epistemic* one, which certainly can be significant, but nevertheless does not undermine the strategy as a whole. In this paper, it will however be argued that the problem is an *ontic* one, that the indeterminacy runs deep and that adherents of means paternalism appear to be laboring under an assumption that can be understood in terms of a *myth of the hidden*: simply taking it for granted that for every individual there is some determinate set of ends that make up their true preferences.

Accepting the ontic indeterminacy of what counts as being in an individual's best interest need however not undermine all forms of means paternalism. In the final section of the paper, it will be argued that there is an alternative approach available, one which arguably makes more sense in a policy-making context anyway, at least for those who are already committed to the distributive policies of existing welfare states. It is an approach that might be dubbed *generic means paternalism*, since it is based in a conception of *primary goods*, the kinds of things that tend to be useful for most of us most of the time.

2 Promoting People's Own Ends

For means paternalism to be viable, we need to understand in what sense the policies suggested by would-be means paternalists are supposed to promote

1 Le Grand and New explicitly distinguish between means-related and ends-related paternalism and reject the latter (2015: 101–104), while Conly suggests that insofar as legislators "are good paternalists, they will try to do what is good for people according to the way people themselves conceive of their good" (2013: 119). Sunstein notes that "for the most part, behavioral economists have not sought to revisit people's ends, and their findings do not support ends paternalism. They have generally emphasized human errors with respect to means, and hence means paternalism is their principal interest and also my main focus" (2014: 63), and Thaler states that "we have no interest in telling people what to do. We want to help them achieve their own goals" (2015: 325).

people's own ends. Now, in some vague sense of *ends* many paternalist policies clearly seek to deliver us goods that most of us want. Default enrollment in saving plans might make us save more money and leave us with a more financially secure retirement. Enacting regulations that lead to a decrease in sodium levels in foods might reduce the risk of heart disease and stroke. Regulating gambling might lead to fewer people ending up in financial ruin. Restricting payday loans can lessen the risk of people ending up with crippling debt burdens. Requiring helmets for bicyclists might lead to fewer head injuries. Mandating that calorie information be clearly presented to consumers might steer us away from unhealthy eating and drinking, as might seeing to it that healthy products are placed more visibly, reducing the size of soda cans or introducing a tax on sugary beverages.

For all of the policies listed above, there are debates to be had about just how effective they are, but in what follows here the focus will be on what they would deliver *if* successful. Even if they are focused on promoting things that most of us want, such as health and wealth, they also involve sacrifices of things that we want as well, albeit often in terms of smaller, more momentary things. It is therefore not unsurprising that a lot of the theoretical discussion has turned on the notion of *preferences*. After all, there can be various things we want in some sense, but which we consistently do not choose since there are other things we want more. If we vastly prefer A to B, then even if B is something that we mildly enjoy, preventing us from choosing A and delivering B to us instead, would certainly seem *prima facie* problematic. Should not then a means paternalism that is to make good on its name involve policies that deliver to us not just what we want, but what we really prefer?

This leaves the means paternalist with a challenge, namely of accounting for relevant preferences, and for the many cases where there would have to be a gap between these preferences and our actual choices.

In considering the approach favored by Thaler and Sunstein, Sugden (2016) has argued that there are two main possible interpretations available here, one in terms of what might be called *latent preferences*, the other in terms of *already existing preferences*. An example of a remark that points towards the first interpretation is when Thaler and Sunstein suggests that well-being ultimately consists in what people would choose if "they had complete information, unlimited cognitive abilities, and no lack of self-control" (2003: 1162). This is a type of position that has been embraced by a number of philosophers, with Henry Sidgwick being an early proponent.[2] As a version of preferentialism about the human good it involves a move that Hausman (2012: 102) has called *preference purification*,

2 "A man's future good on the whole is what he would now desire and seek on the whole if all the consequences of all the different lines of conduct open to him were accurately foreseen and adequately realised in imagination at the present point of time" (Sidgwick 1907: 111–112).

starting with the preferences that we already have, but ultimately considering the preferences that we *would* have. In what follows here this version of ourselves will be referred to as a RICO (a Rational, Informed, and in Control version of Ourselves). While this type of move is relatively common, it should also be clear that it is far from risk-free. By introducing a gap between our current selves and RICOs, we introduce the possibility that the gap is very wide: that what we want and what RICOs would want would be quite different things, and that RICOs are more like some alien beings (see Rosati 1995: 311). Especially for a means paternalist, emphasizing the preferences of RICOs might accordingly risk undermining the very rationale for distinguishing between ends and means paternalism.

The second interpretation is hinted at by Thaler and Sunstein when they say that the goal is "to influence choices in a way that will make choosers better off, *as judged by themselves*" (2008: 5). In fact, the way they reason in terms of a dual-process model might suggest instead that the relevant gap is not one between our present self and a RICO, but between our inner Mr. Spock and our outer Homer Simpson (Thaler and Sunstein 2008: 42). And, certainly, many of the interventions that paternalists tend to propose do concern ambitions or goals that most of us have at times expressed endorsements of: to eat less, to eat healthier, to drink less alcohol, to exercise more, to save more money for future needs and so on. The attractiveness of this interpretation should be obvious: it would seem to remove the threat that affirming means paternalism presupposes making speculative judgments about what our own ultimate preferences *really* are. In Section 4, we will return to a version of this approach, but at least initially, there are at least three reasons why a means paternalist might want to appeal to ideas about what RICOs would prefer.

To begin with, there is a question about how the relevant results in empirical psychology and behavioral economics should be interpreted. In many cases, the very standing of these findings as important results depends on a comparison between how people actually choose and how an agent that chooses in accordance with the tenets of rational-choice theory would choose. There is ample evidence of there being *differences* between the two, but how should these differences be interpreted? The interpretation that means paternalists tend to lean on is that these findings show how we are often irrational, but another possible interpretation is that we are simply different and that the standard model of rationality is misguided if taken as a complete picture of what rationality means for human beings. For instance, Gigerenzer (2015: 379) has argued that the justification of nudging "assumes overly narrow logical norms of rationality." To the extent that we want to understand the choices of Econs (as Thaler and Sunstein refer to them) as having normative authority over our actual choices, we need to bridge the gap between them and us, and it is difficult to see how this can be done if we do not find that the

notion of RICOs makes sense, since what RICOs represent is precisely the meeting point between us and Econs. There is, of course, always the possibility of moving to some alternative understanding of reason and rationality instead,[3] but unless means paternalists are willing to give up on the contrast between us and Econs, which would weaken their case for likening a lot of what we do to acting like a Homer Simpson, they are committed to RICOs being meaningful.

Secondly, in looking at choices between different things that we want in some sense, our occasional cool-hour judgments about what is best hardly settle the question of what we actually *prefer*. Preferences, at least in the sense that is ultimately of interest in this context, can be understood as *total subjective comparative evaluations* (Hausman 2012: 35), which means that if we prefer A over B there is no rational room for further evaluation between having this preference and deciding which of A and B that is the best. In everyday situations we often talk about preferences with respect to partial rankings, say, one's preferences for wines simply by taste, but if what is at stake is something like whether we should eat *tastier but less healthy* or *healthier but less tasty* then we need to know more than just that we want both tasty and healthy, we need to know what we ultimately prefer if there is a choice to be made between the two. Let us assume that for many people their revealed preference is for tasty food, but when asked in a cool hour of reflection they express a preference for healthy food. One possible interpretation of this inconsistency, the one that means paternalists might be inclined to go for, is that the latter preference is our true preference. But is this obvious?[4] It is certainly the preference that is in line with societal norms, but then that might just mean that it is something that we feel expected to say. To be able to come down in favor of the "preference" not steering choice as the true preference, the idea of RICOs as having evaluative authority can play an important role by enabling the argument that what we express in cool hours of reflection is the assessment that is most similar to what RICOs would prefer.

Thirdly, the recent wave of paternalist thinking, to which means paternalists typically belong, is predominately oriented towards paternalism on a policy level,

3 The main alternative here is probably the fast-and-frugal heuristics research program, spearheaded by Gigerenzer (2000, 2008), where the notion of ecological rationality places emphasis on the environments of decision-makers. Another important, but more recent, approach is the argumentative theory of reason developed by Mercier and Sperber (2011, 2017), according to which reason is understood primarily in terms of how it facilitates coordination and cooperation by allowing us to persuade and argue and build more effective groups.

4 Whitman and Rizzo (2015: 410) go as far as saying that there is "a glaring non sequitur at the heart of libertarian paternalism" because of how evidence of internally inconsistent preferences is first used as evidence of irrationality, and then one of those preferences is still selected as the "true" preference.

not specific interventions into the lives of single individuals. This means that the kind of measures that tend to be advocated are often one-size-fits-all interventions. This raises the problematic possibility that the question of how we ourselves understand our own good will differ from person to person (see Rebonato 2012: 160). Maybe some prefer tasty to healthy, others healthy to tasty – or, to the extent that we have more complex preferences, we could differ widely on when and how we prefer tasty and when and how we prefer healthy. If paternalism is not, at least in actual practice, to become an ends paternalism in policy contexts, there would arguably need to be a relatively broad convergence in terms of our ends – otherwise what will be imposed on many of us are means that are not means to our own ends, but someone else's. Of course, the idea of RICOs as determining our true preferences does not necessarily imply that we would converge if we were all in such a state, but at the very least it does hold out a promise of such convergence. After all, there are very many sources of error in reasoning, and accordingly many ways in which we can go wrong; if we remove these sources of error, it seems *prima facie* plausible that there will be less variability in our preferences. The idea of RICOs accordingly has the potential of explaining why one-size-fits-all interventions can make sense.

Now, the idea of RICOs presented here is based in specific formulations by Thaler and Sunstein, but similar conceptions of an idealized version of ourselves have, as already noted, been common in the literature on welfare or well-being for a long time, and the argument here does not hinge on the finer details of how such an ideal is understood. It will accordingly be presumed that for means paternalists it would be a clear advantage if something like RICOs would make sense. Such an ideal can however play at least two roles: *ontic* or *epistemic*. An ontic role means that the preferences of RICOs have a constitutive role in determining wherein our good lies. Facts about RICOs are then truthmakers for statements about what ultimately lies in our best interest. This is the reading of RICOs that is most important, because it is what ensures that when means paternalists talk about how certain interventions are in our best interest, they really can be correct in saying so, even when those interventions run counter to our revealed preferences.

RICOs also playing an epistemic role would mean using this kind of model to actually infer what lies in a concrete person's best interest. In principle, this could involve taking such a person's evaluative and informational states and then deducing which preferences that person would have in her RICO state. In practice, this literal application of the ideal of RICOs would be completely unworkable,[5] yet

5 This is not an issue in descriptive uses of rational-choice models since there we can just stipulate that the relevant agents have certain preferences with respect to the choices under consideration. We do not need fully fleshed-out agents for that kind of modeling. Problems with the messy

at the same time completely giving up on an epistemic role for RICOs would be problematic. It does not help much if there are truths about what RICOs would prefer if these truths are completely unknowable. It is however possible that our inferences about the preferences of RICOs could be made in a looser way, perhaps not showing definitely what the relevant preferences are, but still being enough to give us strong confidence in our assessments about the character of these preferences. We could start by observing which preferences that are typically exhibited by real-world individuals who are more informed, less prone to violating standard principles of rationality and in general more capable of exercising self-control (see Sunstein and Thaler 2006: 256–58). Based on those observations, we could then make inferences about the preferences of RICOs without being fully able to construct complete RICOs for actually existing individuals. A model can be useful even if it does not deliver certainty. Something like this seems to be embraced by Sunstein, who suggests that even "in the absence of reliable evidence about what informed choosers would do … the idea of choosers' informed judgments serves as the lodestar, and it imposes real discipline" (2016: 46). So far, so good, but does the idea of RICOs really make sense, even as a theoretical notion?

3 The Deep Problem of Indeterminacy

The key feature of the above strategy is to cast the problem with applying rational-choice models, and RICOs in particular, to actual choice situations as an *epistemic* problem – not that epistemic problems are insignificant, but arguably they do not fundamentally undermine a given standard of rightness or rationality.[6]

Yet it could also be the case there is a deeper problem with one of the key notions involved in making RICOs tick, namely *preferences*. Indeed, Thaler and Sunstein (2003: 1161) themselves express certain worries about this very notion:

> In many domains, people lack clear, stable, or well ordered preferences. What they choose is strongly influenced by details of the context in which they make their choice, for example default rules, framing effects (that is, the wording of possible options), and starting points. These contextual influences render the very meaning of the term "preferences" unclear.

character of real-life agents instead come later in the process when making inferences about real-world behavior based in results from such idealized models.

6 In consequentialist theory, it is common to distinguish between something being a criterion of rightness and something being a decision procedure (e.g., Bales 1971), the idea then being that even if we cannot in actual practice always tell what is right to do according to the correct criterion of rightness that criterion can still determine what is, in fact, right.

This kind of point is very much in line with how empirical research has shown that when clearly stated preferences are elicited from subjects, then depending on how those preferences have been elicited, they come out differently, even when the methods of elicitation should be normatively equivalent (Grether and Plott 1979; Tversky and Simonson 1993; Slovic 1995). Preferences appear to be partly constructed through how they are elicited. And to some extent this is only to be expected, given that the notion of preferences as total subjective comparative evaluations is already from the start out of line with how human beings make decisions. While we have no problem being subjective, we are rarely (if ever) in a position to make *total* evaluations. Even if we relativize matters to the agent in terms of what counts as relevant considerations,[7] real-world agents are always under time constraints and have to settle for partial evaluations. Normally this is enough to navigate the world; but strictly speaking, having a preference in this everyday sense is not to have a preference in the technical sense.

The obvious response from means paternalists is that this preferential flux mainly concerns momentary preferences, and that this is compatible with our basic values being much more stable. For instance, Conly (2013: 124) suggests that "much of the indeterminacy we see seems to relate to means rather than end. We know where we want to go, but aren't clear on what means are best, and thus are susceptible to the influence of nonrational factors." Since one of the founding ideas behind means paternalism is that people choose the wrong means, the fact that those more specific preferences are indeterminate would not seem to constitute any significant problem. But why assume that the indeterminacy stops there? The bare fact that we might perhaps list a number of basic values that we adhere to, for example, freedom, health, pleasure, companionship and success, does not mean that we have clear preferences between them. Additionally, many of the kinds of behavior that paternalists seek to curb certainly seem to involve at least some pleasure (surely an important good for many of us), so just vaguely gesturing to an unspecific set of broad ends does not establish much in terms of justifying paternalist interventions. What the means paternalist needs is not just that we value health, but that we actually prefer, say, a somewhat longer healthy life to a somewhat shorter life filled with tasty pleasures. What is needed is not just to be able to say that we typically have certain ends, but that there are determinate preference orderings for these ends. And if we understand the issue this way, it is far from clear that indeterminacy is simply a surface phenomenon. Indeed, given the constraints in place on real-life choices it is perfectly understandable why we do not have clear preferences between such larger ends. We rarely, if ever, face

7 On Hausman's account (2012: 2) "a total comparative evaluation takes into account *every* consideration the agent judges to be relevant."

definitive choices between such broad goods; typically we choose between minor instances of them, and under normal circumstances we can have a little of everything. To a large extent, it will be *chance*, in terms of which choice situations with which contingent characteristics that we happen to face and when, that will determine the exact balance between them in our lives.

There is also a further problem here: while many of the more specific choices that we face are between fairly determinate objects, such as *this deep-fried Mars bar here* or *that apple there*, the larger values to which these choices might contribute or be in line with are typically much less so. You could probably find a person who would express an affirmative preference for, say, freedom over companionship; but ask that person about what freedom is or what counts as companionship, or how different aspects of freedom or different forms of companionship might differ in importance, and it will in all likelihood become obvious that while the expressed preference might sound clear, its objects are really far from determinate. In fact, for many of us, were we to encounter a person who had fully worked-out conceptions of matters like freedom and companionship, it is highly doubtful that we would regard that person as an exemplar of wisdom, but perhaps rather the opposite. The reason is that, at least in the world that we inhabit (and given all our limitations), it makes sense to partly make things up as we go along. The only way to have a determinate life plan, and the detailed shape-of-life preferences that come with it, is to neglect the complexities of leading actual human lives. If we do not let our understandings of things that matter be gradually shaped by our experiences, but think of these as something best decided once and for all in a cool hour of reflection, we are fundamentally misunderstanding what it means to lead a human life. Indeed, if we think that one common "misconception is that preferences predate social contexts" (Sunstein and Thaler 2006: 235), we should also recognize that in terms of human design characteristics, this kind of openness to letting things be more precisely determined in context is really not a bug, but a feature.

When it comes to how we, more concretely, navigate our lives, it should be kept in mind that the idea of RICOs does not as such presuppose that we, *qua* actual agents, try to emulate RICOs in our everyday life. As an idealization it could still be useful to us as theorists or would-be paternalist policymakers. The problem, however, is that for RICOs to make sense there still needs to be a path from one's current self to some uniquely determinate RICO state. Otherwise, the problem of indeterminacy is not just *epistemic* but *ontic*, with the idea of RICOs facing a problem of multiple determinability: just as our everyday preferences will partly be constructed through an interaction between us and contingent features of our choice environment, the preferences of RICOs can also take many shapes depending on a range of contingent features. There are two main problems

concerning the process of transformation from one's regular self to a RICO, both of which arguably involve ontic rather than epistemic indeterminacy. One concerns the end state of the process. As pointed out by Grüne-Yanoff (2012: 642), "it is unclear what complete information is, and what unlimited cognitive abilities and self-control means." This is a serious problem in its own right, but not the main point in the present context.[8] The other problem concerns the path to that end state, even if it were clear what the end state ultimately involved.

The main worry is about the input into this process of transformation: since my current state is motivationally and evaluatively indeterminate, there is no authoritatively determinate version of me that can, even in principle, be fed into this process of transformation. Note that the idea here is not that I am in an utter state of indeterminacy, but rather that already before we can put my current self through a process of preference change, my current motivational and evaluative states need to be explicated in terms of a set of preferences that could then undergo preference change. The problem is not that I am in an indeterminate and incomplete state as a physical entity in the world, but that as a person whose motivational and evaluative states are to be expressed in terms of preferences, I am not in a determinate state compatible with the relevant modeling. It is not just that I, in the end, need to be transformed into a different (and better) version of the kind of being that I already am, but that I need to be transformed into a different kind of being already for that process of transformation to begin. The problem is accordingly an *interface problem* between me and the logic of preference change that needs to be applied in order to arrive at what my RICO would prefer, and hence for there to be a determinate answer, even in principle, as to what such a process of transformation would yield.

It might very well be the case that I am in a fully determinate state of being in the sense that I have a set of determinate dispositions to form certain preferences given certain stimuli – but those dispositions will be grounded in a variety of heuristics and biases, which means that any attempt at purifying my preferences will presuppose those limitations in rational abilities that the purified preferences are supposed not to be characterized by. There is accordingly never a pure starting set of motivational states that can be disentangled from the messy motivational and evaluative states characterizing me as an actual agent. I would have to be something much more like a RICO already at the outset in order to be fully transformed into a RICO in the end.[9] As already pointed out, some critics of the idea

8 Sobel (1994) looks at the full information aspect and convincingly shows how problematic it is.
9 One possible objection here is that this argument takes the idea of RICOs too literally; rational-choice models involve idealizations, and as Paul Weirich (2004: 33) suggests, "an idealization's purpose is the truth of a principle employing it, not the truth of a conclusion derivable from it".

of RICOs worry that my RICO will be too alien to me, given that the changes are significant enough, but the point made here is not that the end state, the RICO, would be so different from me in terms of beliefs, values and preferences that it might be questioned whether my individual RICO really is a refined version of *me* specifically, but rather that the whole idea of RICOs are, for want of a better expression, alien to the human condition.

On this reading, the root problem with the underlying psychology assumed by means paternalists is not so much, as Sugden (2018) has suggested (and as the Spock/Homer Simpson metaphor used by Thaler and Sunstein might imply), that they are laboring under an idea about an inner rational agent trapped by an outer psychological shell, but rather that they are laboring under an idea that might be called a *myth of the hidden*, mistaking an ontic problem of indeterminacy for an epistemic problem of determinability by assuming without argument that our entanglement in the world is something that in principle can be abstracted from in a way that leaves us with something pure and true – that the rails are already, so to speak, laid out in a definitive direction. But our entanglement really runs deep. We can undergo change, and some such processes of change could lead us towards states where we are better informed, have sharpened cognitive abilities and more self-control, but there are many different ways in which we can start such processes, and no uniquely privileged way of doing so, and there is at least on the outset no principled reason for believing that all of these paths ultimately lead to the same, albeit hidden, place. Instead, if we start with our subjective motivational sets, in many cases there will simply not be any determinate answer as to what is rational in a maximizing sense, because there simply are no determinate enough ends to be maximized.

All of this is not to say that we can never identify certain things as clearly being in a person's interest and others as decidedly not. For instance, Bernheim (2016), who accepts that we should drop the assumption of there always being some kind of "true preferences" that can somehow be uncovered, suggests that even if

---

Idealization can allow us to articulate certain principles which help us understand, or explain, what it would mean for a decision to be rational, but that need not allow us to identify which decisions count as being rational in concrete situations. The argument here, however, is not against certain principles of rationality as normative standards *when they are applicable*, but rather about how their limited applicability to human choice means that it is not reasonable to see them as constituting a comprehensive standard of good human decision-making. We can still accept, for instance, that *ceteris paribus* if there is a Pareto improvement that can be made, it ought to be made; but accepting such a principle does not actually presuppose accepting rational-choice models. All that these models add is a description of a fictional world where the relevant *ceteris paribus* clauses can be fully satisfied, whereas in the real world such a principle will almost always just have to be a rough guide.

preferences are typically constructed situationally, there might still be cases where we can quite reasonably be said to make mistakes. There are situations that involve what he calls *characterization failures*, that is, where the judgment by an agent is incorrectly informed in terms of the relationship between the alternatives one faces and their outcomes if chosen, and where there is another option that would consistently be selected barring such characterization failures. Take Williams' (1981: 102) classic example of a person who is about to drink from a glass, thinking that it is gin and wanting to drink gin, but where it is really petrol. Even though Williams is a subjectivist about what we have reason to do, he still concludes that this person has reason *not* to drink from the glass. In Bernheim's terms, the preference for gin over petrol is a case where one alternative is unambiguously superior to the other. One never chooses petrol over gin simply because of the finer details of the choice architecture, one only does so when there is a clear-cut characterization failure. No need for an appeal to a RICO here.

But what types of policy might be justified on these grounds? The kinds of things that will tend to be identified as mistakes on this picture will tend to be cases where people do not have good information or fail to comprehend the information available to them, and Bernheim himself seems to favor the kinds of interventions that have come to be known as *boosts* (Hertwig and Grüne-Yanoff 2017), aimed at improving people's competence to make their own choices rather than making those choices for them. It also seems likely that for many choices there will not be any unambiguously superior alternatives, but rather that we will have a variety of desires, where different ones will come out on top depending on the circumstances. This does not invalidate the idea of means paternalism, but it seems highly likely that its scope would be considerably narrowed if one simply focused on addressing clear-cut characterization failures. Thoma (2021), elaborating on Bernheim, suggests that "unless we have clear evidence, with a high burden of proof, of a mistake due to characterization failure, we should presume that actual choice is a good guide to what serves an agent's desires best all-things-considered." Many of the policies that tend to be suggested by means paternalists will probably not pass this test.

Of course, the goal here should be to explore what a reasonable form of means paternalism would look like rather than defending the idea that it has wide-ranged application. However, there is one dimension of the problem of indeterminateness that is not really addressed by focusing on clear-cut mistakes, namely that our motivational states change over time, and that many of the choices we make now have long-term consequences.[10] This is very much a problem already for the idea of

10 Another problem with focusing on revealed preferences, pointed out by Haybron and Alexandrova (2013: 170), is that many of the more principled values that we hold might not really come

RICOs, because in which version of oneself should the RICO that determines what lies in my best interest be based?[11] But it is also a problem that does not disappear simply by rejecting the idea of RICOs. As argued by Hanna (2018: 105–06), there is no obviously right strategy with which a means paternalist can address this issue, for example, *older* does not necessarily mean *wiser*. And unless we are laboring under something like a myth of the hidden, believing in something being there simply because we want it to be there, there is really no reason to believe that there is any straightforward answer, based in my desires or preferences, to what lies in my best interest given that I will typically change over time and that different consequences of my choices will occur at different times in my life. Arguably, these issues raise a question about whether the focus on clear-cut mistakes favored by Bernheim and Thoma narrows things down too much.

We can distinguish between at least two ways in which there can be a lack of unambiguously superior alternatives. One is that there are different choice situations where our desires are fully actualized by alternatives that we face, but where even without any characterization failures being involved we sometimes choose one option, sometimes another (for example, sometimes I drink water, sometimes I drink gin). The other is where the relevant desires are only incompletely actualized by the alternatives. For instance, I might enjoy salty snacks, but I also do not want to have a heart attack. On every given occasion where I eat salty snacks, however, the effect of my doing so on that particular occasion will be negligible in terms of the risk for having a heart attack. This means that one relevant long-term effect is really never on the table as a consequence of any particular choice. We are often faced with something like an intrapersonal version of the tragedy of the commons: one cigarette, one bowl of salty snacks or one jumbo-sized soda will not make any difference, but consistently choosing such alternatives one time after another, risks cumulatively realizing effects that are clearly bad given my own desires. Since those cumulative effects never fully come into play in particular choice situations, relying on actual choice as a guide to what best serves my actual desires seems problematic.

Even when there are available alternatives where I can more directly opt for a certain long-term effect, such as when signing on to a savings plan or an insurance policy that will thereafter operate automatically, one might still worry about whether present-time biases might not affect my choices in problematic ways, even

through in our everyday choices, since they might not be at stake in them; taking actual choice as a guide can then come with the risk of not recognizing such values. There is no room here to discuss this concern in any detail, but the pluralistic approach to the good that they propose is arguably at least somewhat in line with the kind of framework suggested in Section 4.

11 For an overview of such issues, see Bykvist (2006).

if there are no characterization failures involved. My future self is helpless with respect to what my present self might come up with. This is a fundamental asymmetry that is not really addressed by a focus on characterization failures, however relevant those are in many other cases. We are not just agents making choices at different moments and in different circumstances, but persons leading lives where the effects of our choices are often cumulative and can take place at very different points in time. To the extent that a more wide-ranging form of means paternalism is warranted, even in the face of the problem of indeterminacy, the strongest reason for taking such an approach is arguably to address these kinds of issues.

4 Already Existing Ends and the Idea of Primary Goods

Since we have rejected the latent-preferences approach, then at least if we do not want to take some kind of perfectionist route, what remains is some version of an actual-preferences approach, although given the arguments above that kind of approach might perhaps best be articulated not in terms of preferences, but in terms of our own ends or our own desires instead. As noted in Section 2, however, this option will come with certain costs, and the question is how these costs can be minimized. One cost was that without RICOs as a contrast it might be more difficult to make the case that certain behaviors are irrational or constitute mistakes. Another was that for many choices we will have to accept that there might not be an answer as to what is *best* for a given individual. The third problem was that a possible argument for why certain one-size-fits-all polices can be warranted will be unavailable.

How serious are these issues? This partly depends on our overarching framework, and the conclusion of Section 3 pointed us towards taking a life-cycle perspective on people, where many of the choices they make in the present also affect what will happen to them and how they will fare later in life. This kind of perspective is not about arguing that certain actions are irrational given the relevant desires in the moment of choice, but rather about looking at what is reasonable from a life-cycle perspective. Taking this kind of perspective is very much in line with how contemporary welfare states tend to function, where a significant (perhaps even predominant) part of their redistributive effects occur *within lives* rather than between distinct individuals (Bergh 2005). This suggests that when looking at choices from this perspective, it will be reasonable to do so within the broader scope of considering what a well-functioning welfare state

looks like. Of course, on a theoretical level, a society built on libertarian principles, with a minimum of government interventions into our lives, remains a possibility. But in actual practice, governments in all advanced economies already take an active role in shaping the rules of the game that govern how we are able to pursue our own personal ends.[12]

In discussing paternalist policy-making it is accordingly important to keep in mind that in reality it is not just a possible extra set of policies that we are considering, but that paternalism is already built into the functioning of existing welfare states. In considering this fact, Ben-Porath (2010: 19) even goes as far as to say that "paternalism is an inevitable and indispensable aspect of social relations and of democratic policy making."[13] Of course, there is still a question about just how much of paternalist policy-making we should have, and which exact policies, but on the whole we are facing a question of *how* rather than *whether*. If we have this framing of the matter, that is, as one about how to best design a welfare state, the worry about some paternalist policies being one-size-fits-all solutions will arguably have less bite. Existing welfare schemes typically work by offering particular suites of goods or opportunities, for example, access to health care, education, housing, and where we are not free to trade some quantity of one of these in order to gain more of another. This certainly does not mean that one-size-fits-all paternalist policies are unproblematic, and there can be ways in which welfare-state policies in general can be designed so that individuals can have meaningful choice; the point is just that in debating different paternalist policies, it should be remembered that in being one-size-fits-all, they do not introduce a new and foreign element into what we already do.

If we turn instead to the first worry, about whether losing the contrast with RICOs undermines some arguments about particular behaviors being problematic, then one thing involved in taking a life-cycle perspective is that specific choices will not be put into question by a contrast to some RICO, but by thinking in terms of how current desires might not be representative of what people want at different times in their lives. In the literature, the main existing account that most clearly exemplifies this kind of approach is arguably the Rawlsian idea of *primary goods*, that is, a list of the kinds of things that we can reliably assume that people would

12 It should be recognized that there are ways in which the welfare state could be designed to minimize intervention, universal basic income being the main alternative (e.g., Van Parijs 1995). Existing welfare states are however not designed like that and still enjoy broad political support.
13 The *structured paternalism* suggested by Ben-Porath is however more strongly egalitarian than what is backed up by the life-cycle focus suggested in the present paper, and her argument also takes into consideration the effects that policy-making will have on civic equality and the capacity of individuals to take part in democratic decision-making as equals.

typically want more of rather than less.[14] Primary goods are identified precisely on account of being *generically important* goods, and if we are subjectivists we should determine these by starting with people's actually existing desires or preferences. This type of analysis will presumably still be one of instrumental rationality: if *this* and *that* is what most people ultimately strive for, then *these things* will be generically valuable as means to those ends. The most substantive presumption typically made in this kind of analysis is precisely one of temporal neutrality (or at the very least only modest temporal discounting of goods), which then has to be coupled with a recognition of how what happens at early stages of a life influences later stages in a highly asymmetric way.[15] Exactly what stance we should take on this can be debated, but the important point here is that taking this kind of life-cycle perspective does not as such presuppose the viability of RICOs, it is primarily based in ideas about personal identity, that we are the same persons throughout our lives and that there are no independent reasons for privileging any particular stages of those lives.

The ontic problem of indeterminacy arose from the need to be highly specific about what RICOs would prefer, but it does not stand in the way of identifying generic goods. Things like health, wealth and income are all highly useful in pursuing different life projects, and things like liberty or autonomy are typically important presuppositions for being someone who pursues his or her conception of the good. There are, of course, still questions about which polices have the *best* outcome on the whole, and these can certainly be difficult as well. On the approach suggested here, this is however primarily a question to be addressed on a population level, where some of the indeterminacies that obtain with respect to each person as an individual will disappear; for example, health risks are much more clearly ascertainable on a population level than for specific individuals. It can be

14 In identifying relevant goods to be promoted by government policies, another option in the literature is the *capabilities approach*, advocated by theorists like Sen (1999) and Nussbaum (2006). It would however seem to move us at least somewhat closer to a form of perfectionism compared with a primary-goods approach. For discussions of the merits of these two approaches, see Brighouse and Robeyns (2010). A third possibility would be to focus on opportunities, as suggested by Sugden (2018), while Hausman (2019) suggests that even if only coarse-grained judgments about what is better for people are possible, one might still use people's actual preferences as indicators when making those judgments.

15 Rawls' own way of framing this matter involves assuming that "a person's good is determined by what is for him the most rational long-term plan of life given reasonably favorable circumstances" (Rawls 1999: 79). The idea that there is such a thing as *the most rational long-term plan of life* seems like yet another instance of the myth of the hidden. The idea of primary goods does however not need this kind of assumption, rather one might focus on identifying goods that are continuously valuable as means for various different life projects throughout people's lives.

unclear if policy A is in *my* best interest, but still be clear that it is in *our* best interest *qua* members of a given society with scarce resources at our disposal.

Additionally, even if we will not always be able to say what is best, the relevant cost–benefit analyses are types of analyses that we are already making, whether explicitly or implicitly, with respect to welfare programs promoting the relevant goods. The fact that the interventions are paternalistic does not add any further complication in this sense. Given a primary-goods approach, the main values typically counting for interventions will be health, wealth and income. The main values typically counting against interventions will be liberty and autonomy.[16] One value that is conspicuously absent from standard lists of primary goods, however, is pleasure or enjoyment. This is a good that we as individuals might choose to pursue, and might even regard as an important final end, but where government policies promoting primary goods will typically merely facilitate such pursuits. In considering promoting healthy consumption of foods and drinks, this means that on this kind of analysis, the balance of values to be struck is not between short-term pleasures and long-term health consequences, but mainly between promoting positive long-term health consequences and avoiding infringements on our liberties or our autonomy.

Three main types of tools that can be used in paternalist policymaking are to (i) simply remove certain options from the choice menus that we face, (ii) increase the cost of choosing those options, or (iii) to tinker with the choice architecture so that we will tend to go for certain options just because of how we happen to function as decision-makers (nudges). If we understand our liberties and autonomy as tied to having options which we might want being available to us, then (i) is typically a more serious type of infringement. But certain interventions, like banning the use of trans fats, will arguably not amount to a serious infringement in terms of the options available to us (maybe somewhat shorter shelf-life for certain products) and could be relatively unproblematic. Proponents of (iii) understand these as the least serious infringements and, to the extent that we share this view,[17] the main question will be how effective such measures are, for example, to what extent posting calorie information really leads to healthier patterns of eating and drinking. When it comes to (ii), different forms of taxation can obviously be used to

16 The more precise ways in which we understand autonomy and liberty will however be significant for what these trade-offs look like. For instance, when it comes to things like drug use and gambling, addiction or addiction-like behavior can be seen as undermining people's autonomy, and in that case paternalist measures can even be autonomy-promoting (even if they might still infringe our liberties somewhat).

17 If we understand autonomy as involving not being manipulated, there might be a case against at least certain forms of nudges as being manipulative (Conly 2013: 30). But this is a further discussion.

steer behavior, and this is certainly an area where a number of interventions are already in place. As indicated by the notion of *sin taxes*, some of these measures might at times be moralistic in character, but as long as measures are clearly geared towards promoting population health, then that end is clearly a legitimate goal from a primary-goods perspective. There are however well-known problems with taxation as a steering mechanism, such as the risk of creating incentives for smuggling and black markets; so to the extent that health can be promoted through means like nudges, there can be good reason to opt for such measures instead. Means paternalism, and this goes for generic means paternalism as well, should not be about moralism, but about using the most efficient means for promoting relevant goods such as population health.

When looking at different measures within the context of an existing welfare state it is important to keep in mind that many of the relevant systems really are communicating vessels. This also means that in making some of these cost–benefit analyses, we will be comparing paternalist policies with redistributive policies seeking to achieve similar goals, although with different means. For instance, it could very well be possible that a certain level of health promotion could be achieved through relatively innocuous paternalist interventions, such as nudges, which then in the long run might lessen the need for publicly financed health care – by preventing health impairments rather than trying to fix them after they have occurred. In one case, a limitation in our freedom or autonomy might then come in the shape of steering our choices in certain ways, in the alternative such a limitation might come in the form of higher taxes. It seems reasonable that what we ultimately would like is to minimize limitations to our freedom and autonomy, and these policies can then at least partly be compared with each other on that metric.

This is not to say that governments automatically have a right to make paternalist interventions in order to safeguard the already ongoing taxpayer investment in, say, my health, but merely that there is at least a *prima facie* reason for such interventions: my individual health is, to some extent, a matter of public concern. As long as we are making policy within the context of a welfare state, there cannot be any strict separation between paternalist policies and classic welfare programs – they will often be different means to the same end. Of course, as already noted, this type of *prima facie* reason then has to be balanced against other reasons and one should not expect there to be any precise algorithm for striking the relevant balances.[18] But in making these judgments, it would at the very least seem reasonable that to the extent that certain behaviors of people

18 Kniess (2021), who also favors a primary-goods approach to reasoning about possible paternalist interventions, suggests that deciding how these balances should be struck more exactly should be decided at least partly through democratic mechanisms, which sounds sensible.

impact negatively on matters concerning certain primary goods for themselves *and* those behaviors also tend to run counter to a balanced temporal distribution of primary goods across people's own lives, these behaviors become especially attractive as targets for policy interventions.

5 Concluding Remarks

While there are reasons for why one might want to be able to appeal to RICOs as defining wherein people's good lies, it is an idea that is ultimately untenable because of the deep problem of indeterminacy. It has however been argued here that a generic version of means paternalism can still be maintained as an integrated part of an understanding of the welfare state as organized around providing its citizens with generic goods that they can use in pursuing their own individual ends or life goals. What we end up with, then, is an approach to paternalist interventions that is still non-perfectionist in that it emphasizes the importance of people leading their own lives. It is certainly an approach that will be less attractive to those truly committed to the *libertarian* part of the libertarian paternalism proposed by Thaler and Sunstein; but, on the other hand, however influential the idea of using nudges has been, governments that have introduced such policies have still tended to remain committed to at least modest welfare-state regimes. At least for the would-be paternalist who is already on board with some such welfare-state regime being in place, the indeterminacy of the individual's own good does not present a strong objection to a moderate level of paternalist interventions or to being a means paternalist.

Acknowledgments: Work on this paper was enabled by the Research Program on Science and Proven Experience (VBE), funded by the Swedish Foundation for the Humanities and the Social Sciences (Riksbankens Jubileumsfond). Previous versions of it have been presented at the Higher Seminar in Practical Philosophy and the Seminar in Medical Ethics, both at Lund University, and the author is very grateful to the participants there for their helpful comments, as well as to the reviewers for this journal for their constructive and useful suggestions.

References

Bales, E. 1971. "Act-utilitarianism: Account of Right-making Characteristics or Decision-Making Procedures?" *American Philosophical Quarterly* 8: 257–65.

Ben-Porath, S. R. 2010. *Tough Choices: Structured Paternalism and the Landscape of Choice*. Princeton: Princeton University Press.

Bergh, A. 2005. "On Inter- and Intra-individual Redistribution of the Welfare State." *Social Science Quarterly* 86: 984–95.
Bernheim, B. D. 2016. "The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics." *Journal of Benefit-Cost Analysis* 7: 12–68.
Brighouse, H., and I. Robeyns, eds. 2010. *Measuring Justice: Primary Goods and Capabilities*. Cambridge: Cambridge University Press.
Bykvist, K. 2006. "Prudence for Changing Selves." *Utilitas* 18: 264–83.
Conly, S. 2013. *Against Autonomy*. Cambridge: Cambridge University Press.
Gigerenzer, G. 2000. *Adaptive Thinking: Rationality in the Real World*. New York: Oxford University Press.
Gigerenzer, G. 2008. *How People Cope with Uncertainty*. New York: Oxford University Press.
Gigerenzer, G. 2015. "On the Supposed Evidence for Libertarian Paternalism." *Review of Philosophy and Psychology* 6: 361–83.
Grether, D., and C. Plott. 1979. "Economic Theory of Choice and the Preference Reversal Phenomenon." *American Economic Review* 69: 623–38.
Grüne-Yanoff, T. 2012. "Old Wine in New Casks: Libertarian Paternalism Still Violates Liberal Principles." *Social Choice and Welfare* 38: 635–45.
Hanna, J. 2018. *In Our Best Interest: A Defense of Paternalism*. New York: Oxford University Press.
Hausman, D. 2012. *Preference, Value, Choice, and Welfare*. Cambridge: Cambridge University Press.
Hausman, D. 2019. "Enhancing Welfare without a Theory of Welfare." *Behavioral Public Policy*, https://doi.org/10.1017/bpp.2019.34.
Haybron, D. M., and A. Alexandrova. 2013. "Paternalism in Economics." In *Paternalism: Theory and Practice*, edited by Christian, C., and W. Michael, 157–77. Cambridge: Cambridge University Press.
Hertwig, R., and T. Grüne-Yanoff. 2017. "Nudging and Boosting: Steering or Empowering Good Decisions." *Perspectives on Psychological Science* 12: 973–86.
Kniess, J. 2021. "Libertarian Paternalism and the Problem of Preference Architecture." *British Journal of Political Science*, https://doi.org/10.1017/S0007123420000630.
Le Grand, J., and B. New. 2015. *Government Paternalism*. Princeton: Princeton University Press.
Mercier, H., and D. Sperber. 2011. "Argumentation: Its Adaptiveness and Efficacy." *Behavioral and Brain Sciences* 34: 94–111.
Mercier, H., and D. Sperber. 2017. *The Enigma of Reason: A New Theory of Human Understanding*. Cambridge: Harvard University Press.
Nussbaum, M. 2006. *Frontiers of Justice: Disability, Nationality, Species Membership*. Cambridge: Harvard University Press.
Pickett, B. 2019. "The New Paternalists." *Polity* 50: 300–29.
Rawls, J. 1999. *A Theory of Justice, rev.* ed. Cambridge: Harvard University Press.
Rebonato, R. 2012. *Taking Liberties: A Critical Examination of Libertarian Paternalism*. New York: Palgrave Macmillan.
Rosati, C. 1995. "Persons, Perspectives, and Full Information Accounts of the Good." *Ethics* 105: 296–325.
Sen, A. 1999. *Development as Freedom*. Oxford: Oxford University Press.
Sidgwick, H. 1907. *The Methods of Ethics*, 7th ed. London: Macmillan.
Slovic, P. 1995. "The Construction of Preference." *American Psychologist* 50: 364–71.
Sobel, D. 1994. "Full Information Accounts of Well-being." *Ethics* 104: 784–810.

Sugden, R. 2016. "Do people Really Want to be Nudged towards Healthy Lifestyles?" *International Review of Economics* 64: 113–23.
Sugden, R. 2018. *The Community of Advantage*. Oxford: Oxford University Press.
Sunstein, C. 2016. *The Ethics of Influence*. Cambridge: Cambridge University Press.
Sunstein, C., and R. Thaler. 2006. "Preferences, Paternalism, and Liberty." *Royal Institute of Philosophy Supplement* 59: 233–64.
Thaler, R. 2015. *Misbehaving*. New York: W. W. Norton.
Thaler, R., and C. Sunstein. 2003. "Libertarian Paternalism Is Not an Oxymoron." *The University of Chicago Law Review* 70: 1159–202.
Thaler, R., and C. Sunstein. 2008. *Nudge*. Yale: Yale University Press.
Thoma, J. 2021. "On the Possibility of an Anti-paternalist Behavioural Welfare Economics." *Journal of Economic Methodology*, https://doi.org/10.1080/1350178X.2021.1972128.
Tversky, A., and I. Simonson. 1993. "Context-dependent Preferences." *Management Science* 39: 1179–89.
Van Parijs, P. 1995. *Real Freedom for All: What (If Anything) Can Justify Capitalism?* Oxford: Clarendon Press.
Weirich, P. 2004. *Realistic Decision Theory: Rules for Nonideal Agents in Nonideal Circumstances*. New York: Oxford University Press.
Whitman, D., and M. Rizzo. 2015. "The Problematic Welfare Standards of Behavioral Paternalism." *Review of Philosophy and Psychology* 6: 409–25.
Williams, B. 1981. *Moral Luck*. Cambridge: Cambridge University Press.