Res Publica (2019) 25:167–185
https://doi.org/10.1007/s11158-018-9393-6

Regulating Compensatory Paternalism

Johan Brännmark[1]

Published online: 31 January 2018

Abstract Some recent arguments for paternalist government interventions have been based in empirical results in psychology and behavioral economics that would seem to show that adult human beings are far removed from the ideals of rationality presupposed by much of philosophical and economic theory. In this paper it is argued that we need to move to a different conception of human decision-making competence than the one that lies behind that common line of philosophical and economic thinking, and which actually still lies in the background of some of these recent approaches to paternalist interventions. An alternative picture of human decision-making competence is outlined and four criteria for identifying areas where paternalist interventions have a basic moral and political legitimacy are then identified on the basis of this picture.

Keywords Paternalism · Policy-making · Rationality · Decision-making · Preferences

Introduction

A growing number of writers are currently arguing for a variety of paternalist interventions. Thaler and Sunstein (2003, 2008), who could perhaps be seen as the originators of this revival of interest in government paternalism, have put forward nudges as a way of effectively making choices for people but still leaving them with

✉ Johan Brännmark
johan.brannmark@mah.se

[1] Department of Global Political Studies, Malmö University, Malmö, Sweden

the freedom to do otherwise,[1] while writers like Conly (2013) and Le Grand and New (2015) open up for stronger forms of interventions as well. There are many differences between these writers, but one thing that tends to unite them is that much of their reasoning is based in contemporary empirical psychology, especially the heuristics-and-biases literature and the distinction between System 1 and System 2 decision-making. This is a literature which points towards a more complex conception of human agency than the one that has been common in much of economics and significant parts of philosophy, namely that of a utility maximizer with a complete and ordered set of preferences, and for whom rational choice consists in choosing the best means to fulfill these preferences.[2] The psychological literature certainly puts into question the descriptive accuracy of this way of modeling human agency, but one might also then raise questions about whether it should not make us reconsider how we think about government paternalism as well. If irrationality in human choice turns out to be not just occasional but endemic, should we not be more open to government effectively making certain choices for us, as something that is needed in order to compensate for prevalent problematic tendencies in our decision-making?

But while adherents of what might be called *compensatory paternalism* tend to question the traditional picture of rational agency as a descriptive model, they also have a tendency to still rely on it, or something very much like it, as a yardstick both of what human decision-making competence is about and of what our good ultimately consists in, namely satisfaction of the preferences we would have if fully informed and rational (which is certainly not the only conception of well-being out there, but a very common one, especially in the area of overlap between economics and philosophy). There is clearly no logical contradiction in maintaining these ideas as yardsticks, since they are then ideals rather than descriptions, but questions can clearly be raised about their aptness as ideals, and the principal argument in this paper will be that while compensatory paternalists are right in thinking that empirical psychology gives us reason not to trust that human beings always know best how to run their lives, this literature can also be taken to point towards a more far-reaching rethinking of how we understand human agency and human well-being.

There will be three main steps to the argument. First, we will look at the problems contemporary empirical psychology poses for standard ideas about human agency. Second, at how human decision-making competence should be understood in the light of these empirical results. Finally, four criteria for identifying suitable areas for paternalist interventions will be articulated. It should be said that while the argument is partly critical, the paper is mainly an attempt to make a

[1] There are arguments to be had about to what extent nudges simply are about there being one choice architecture in place rather than another (recognizing that there is always some such choice architecture in place anyway) or whether nudges are manipulative (White 2013). Sunstein (2016, Chapter 5) has argued that, at least on a reasonable understanding of manipulation, most nudges do not count as manipulative; but in the present context the important point is simply that nudges, at least when working as intended, will involve government effectively ensuring that most of us make certain choices rather than others.

[2] It should be noted that within philosophy one often finds similar but less formalized Humean ideas about practical rationality as instrumental rationality instead, as well as more clearly dissimilar conceptions of practical rationality. Many of these accounts of practical rationality might also be put into question by contemporary empirical psychology, but this is not an argument that will be made here.

positive contribution in terms of how to think about human decision-making competence and how to assess the moral and political legitimacy of paternalist policy measures.

Understanding Human Decision-Making

Whether framed in terms of restricting our autonomy or not treating us as adults, policies which involve government effectively making choices for us are often seen as highly problematic. But there are also obviously some human beings, children, with respect to whom it is not as seen as especially problematic that other persons, parents, make many choices for them. Presumably, this is because there is something about most adult human beings that warrants decisions being left to us, both formally and effectively. If we look at the philosophical tradition, there has historically been a strong link between emphasizing the value of autonomy in some form[3] and emphasizing rationality as a characteristic that sets adult human beings apart from other agents, and rational-choice models for understanding human decision-making have been prevalent in philosophy as well as economics. Given an emphasis on rationality, there might be two main reasons for leaving decisions to us. One is that it will involve an exercise of our faculties that can be valuable in itself, and the most straightforward way to make sense of that idea is probably to see rational decision-making as *the* distinctive feature of human life, perhaps even as a form of realization of our potential as human beings.[4] The other reason is that if we are left to our own devices, we will because of our rationality and self-knowledge be able to look after our own interests in a way that no one else could.[5] Both of these reasons will be considerably stronger if we are, at least on the whole, rational decision-makers in actual practice and not just in theory.

Of course, the bare fact that two themes can be found together does not mean that one necessarily presupposes the other. Since you cannot derive an 'ought' from an 'is', the value of autonomy and exercising choice is logically distinct from particular conceptions of human agency. Normative ideals should, however, still be more than merely possible to accept without being logically inconsistent. There is also a question of relevance: normative ideals should be pertinent to the choices we face and their degree of relevance will always have to be judged in terms of their fit with that part of reality which they are supposed to be ideals for. This degree of fit will partly be determined on empirical grounds. So while values and facts about human behavior are logically distinct, one can hardly deny that certain ideas form partnerships where the included parties make each other look more attractive by standing next to each other. If the glory of one partner fades, however, one should perhaps then reasonably take a second look at the other partner. While many of the

[3] While there is a very large philosophical literature on how to best understand autonomy, there does not seem to be any consensus forthcoming on just how to best define it (Anderson 2014), so the argument here will not be framed in terms of any particular notion of autonomy.

[4] An idea that can be traced back at least to Aristotle (1999).

[5] This is of course Mill's (1989) main argument.

core ideas about human agency that we find in contemporary moral and political theory are recognizably similar to ideas formulated already during the 17th and 18th centuries, developments in empirical research on human decision-making over the last 40 years or so have been considerably more dramatic, arguably dramatic enough to warrant updates in our moral and political ideals as well, at least if these are to be relevant ideals for actual human beings and not merely for philosophical constructs.

There are many different variations of the standard picture of human agency that we find in the literature, either explicitly in the foreground or implicitly in the background, and it is impossible to cover all variations here, but here are three common ideas about key components of human decision-making:

1. *Motivational:* individuals have (or are in principle able to achieve) clear, ordered, and complete sets of preferences, which are the motivational states ultimately relevant for action.
2. *Informational:* having complete information about something is simply a matter of amassing or pooling all the relevant and true (or likely to be true) bits of information pertaining to that something.
3. *Procedural:* being rational is about being governed by certain general principles of handling information, structuring choices, and making expected utility calculations.

All three of the above components have, however, turned out to be problematic. To begin with, depending on how one attempts to elicit our preferences we make different choices, even when the methods of elicitation should be normatively equivalent (Grether and Plott 1979; Tversky and Simonson 1993; Slovic 1995). We have preference reversals: given one method of elicitation a subject will choose A over B, given another B over A. Arguably, the most reasonable interpretation of these results is that at least for many preferences, we partly make things up as we go along rather than drawing on pre-existing settled preferences. Our preferences tend to be context-dependent and circumstantial factors accordingly play a significant role in determining what we choose.

When it comes to handling information, we do not just piece together information by pooling it, we structure it into pictures and stories. This means that the sequencing of how information is provided will cause order effects (Hogarth and Einhorn 1992) and new information will usually be interpreted to support already existing pictures/stories; we have a strong tendency to confirmation bias (Wason 1960). Additionally, beliefs related to self-assessment often seem to be more about managing one's self-esteem than searching for the truth, so the line between the evaluative and the cognitive is often breached. For instance, in one study of American drivers, 93% of them thought that they were better than the median driver (Svenson 1981) and in a study of professional fund managers 74% believed that they were above average in their job performance while 26% believed that they were average (Montier 2007). When it comes to the likelihood of external events happening to us rather than to others we exhibit similar patterns in that we tend to overestimate the likelihood that good things will happen to us rather than to others

and underestimate the likelihood that bad things will happen to us rather than to others (Weinstein 1980).

Finally, when it comes to principles of reasoning, the traditional model has a picture of decision-making competence that is of a generic/universal kind, that we are governed by certain principles. This is compatible with being, say, risk-averse—but then we should be risk-averse in a principled way. In reality things are not that smooth. For instance, Tversky and Kahneman (1979) found that there is a *certainty effect*, so that people tend to underweight outcomes that are probable compared to those that are fairly certain. This causes people to be risk-averse when they can have a certain gain, but to be risk-seeking when they would otherwise face a certain loss. This also has the effect that depending on how choices and probabilities are presented, e.g. whether something is presented as involving a 80% chance of survival or 20% chance of death, we tend to choose differently. In recent years neuropsychologists have also found that when it comes to abstract objects like numbers, the mental number line gradually gets more compressed (Longoa and Lourenco 2007; Merten and Nieder 2009), i.e. our ability to meaningfully differentiate between numbers that lie close to each other decreases as the magnitude of the numbers increases. The difference between 10 and 20 might be very notable, whereas 110 and 120 might be seen as almost on a par (which means that what is the same marginal gain will be interpreted differently). This is arguably related to the phenomenon of so-called 'psychophysical numbing' (Fetherstonhaugh et al. 1997; Slovic 2007), where we are disproportionally willing to spend money and effort to gain effects that lie at the lower end of what can be accomplished than we are at the higher end, even when the effect on the margin might be the same, e.g. saving ten additional lives. We also tend to suffer from temporal myopia, where future gains and losses are discounted in a way that seems to involve dynamic inconsistencies (Strotz 1955–56; Thaler 1981; Frederick et al. 2002; Benhabib et al. 2010).

There is much more to be said about each of these points and it should be acknowledged that the interpretation of every individual result is certainly contestable (and often already contested). Alternative explanations can always be suggested in order to dismiss a specific result, but for present purposes the most important question is what kind of general picture emerges from these findings. Do they indicate that we now and then fall short of the ideal encapsulated in rational-choice models or is there a more fundamental implication, namely that as agents we function in a distinctly different way? Many psychologists and cognitive scientists have suggested that the latter is the case. More specifically, it is commonly thought that higher-order human thinking can be divided into two main types of processes.[6] First we have what is often called System 1 thinking, which is fast, automatic, and non-conscious. The links made between different items tend to be associative and the items themselves are often metaphors, images, and narratives. Reactions tend to be rooted in affect and gut feelings, often related to previous experiences. Then we

[6] While many psychologists talk about these two kinds of processes in terms of 'systems', the important distinction really is between two kinds of processes or two kinds of thinking (the extent to which these actually form discrete systems is a different matter and one which there is no need to discuss here).

have System 2 thinking: slow, deliberate, and conscious. The connections made between items here tend to be logical and the items themselves are abstract symbols, words, and numbers. There is a conscious assessment and weighing of the support that exists for different beliefs and actions. To put it briefly: System 2 is more or less the kind of thinking assumed by models of rational choice and we do have this capacity. However, compared with System 1 the role it plays in determining our decisions in everyday life is marginal, simply because its use takes so much time and effort and it requires so much information to work on.

Just how entrenched the standard picture is can be illustrated by how it ultimately shows up even in accounts that are built on having problematized it, albeit as a yardstick rather than a description. For instance, Richard Thaler and Cass Sunstein argue in favor of paternalist interventions in the form of nudges based in this kind of empirical research. The point of these interventions is, however, still understood as 'to influence choices in a way that will make choosers better off, *as judged by themselves*' (2008, p. 9) and well-being consists in what people would choose if 'they had complete information, unlimited cognitive abilities, and no lack of self-control' (2003, p. 1162). Of course, Thaler and Sunstein are not actually trying to elicit this set of preferences in anyone, but if as already pointed out, the exact methods of elicitation that are used when actually trying to elicit preferences tend to partly construct these preferences, then it seems highly doubtful that there is a single definitive state of a person's preferences that would appear if you just add complete information and superhuman cognitive abilities (*cf.* Sugden 2008). One could possibly stipulate that this kind of cognitive transformation would be instant and that there would be no need to sequence it (since there would be several ways of sequencing it and these could potentially yield different results), but then there is a question about to what extent this is a workable ideal in actual practice. While we might try to approximately emulate sequential transformations, an instant wholesale transformation is difficult to grasp. Sarah Conly recognizes that the fact that the psychological literature points to indeterminacy in our preferences poses a problem to some extent, but suggests that a great deal of the indeterminacy concerns means rather than ends, and since our well-being ultimately is based in our preferences or desires about ends, indeterminacy is not that much of a problem (2013, pp. 123–125). Similarly, Julian Le Grand and Bill New distinguish between means-related and ends-related paternalism and reject the latter (2015, pp. 101–104)—but this distinction, again, presupposes that means and ends can be disentangled from each other, and if our preferences are often indeterminate this cannot be taken for granted.

Need we really interpret the empirical data as pointing to significant indeterminacy in our preferences? Could it not just be indicative of the dynamic character of preference formation? If so, while constantly changing, our preferences could still be determinate. Irrespective of which interpretation you go for, there is a need to move beyond the data in formulating general models here. The guiding idea in this paper is that if we look at human beings as decision-making systems, to form determinate preferences will always come at a cost in terms of time and energy. For such a system it accordingly makes sense to have preference orders that are incomplete but *settled enough* to function well in everyday life. Maybe we are often

faced with a choice between A and B; then we need a preference to govern that choice. Yet while it is also possible that we will some day encounter a choice between A and C we need not already have a preference for that choice, nor need we have preferences for all the kinds of variations that there might be of A and B. We can partly make things up as we go along and still function well in everyday life. When it comes to hard choices we might even tend not to think about them in advance precisely because they are hard. Take the following example: a Jehova's Witness is in need of a life-saving blood transfusion. She wants to live and be with her family. But she also wants to respect her religious beliefs. When her physician meets with her alone, she prefers the blood transfusion, but when the elders are in the room she rejects it.[7] In this type of case it could be that what she really prefers is the blood transfusion, but it seems false to think that the preferences that we express when alone are always our true preferences; it could also be that in the company of her elders she feels more secure and is able to overcome fears that she wants to overcome. But it could also simply be the case that there is no true preference there and that different ways of eliciting a preference will result in different preferences being formed. This is not to say that there are no cases where we can clearly identify certain preferences as unauthentic; the point is just that we have no reason to think that there is always an answer, even in principle, to the question of which preference is the true preference.

A partial and gappy system of preferences will often be compatible with not having decided whether we want a certain thing as a means or an end (or perhaps both). We tend to know roughly what we want, but it would be a waste of cognitive resources to form ideas about what we want in all conceivable situations and under all conceivable permutations. Of course, knowing roughly what we want is enough for there to be certain core values that are fairly constant, but these will often be very general and abstract, which means that we might not always know what counts as falling under those values. For instance, a lot of people might value freedom and will be able to give some paradigmatic examples of what it means to be free, but might still be uncertain about other examples. When it comes to many concrete applications of our core values we mainly seem to be muddling through, as indicated by phenomena such as moral dumbfounding[8] and choice blindness,[9] rather than being governed by clear conceptions of what we value. So even if we do have ends, they appear to be far too indeterminate for our thinking about paternalism to rely on an idea about means-ends-structured preference orders. That idea is really just a leftover from the old standard picture. In everyday life we do not need

[7] This example comes from Tännsjö (2015, pp. 716–717); the point he makes in relation to it is about how practices of dealing with such cases should be based in what works best in general rather than fine assessments about authentic preferences in individual cases.

[8] Where strong moral convictions about particular cases are coupled with significant difficulties in accounting for those convictions in terms of justifying reasons, see Haidt (2001). The reasons that we provide when asked for them look more like post hoc rationalizations than something really driving our moral judgments.

[9] Which involves situations where we are told to make a stand on a certain matter and when we are later on confronted with evidence that we actually took the opposite stand, there is a strong tendency to start providing reasons for this latter stand instead, see Hall et al. (2012).

strongly determinate ends and so there is no reason to assume that we have them. And even if we had strongly determinate preferences, but where these were constantly changing, it would still be difficult to see how such shifting determinate preferences would be of much help in public policy matters. The fact that our preferences change also creates a problem of deciding at which point in time the state of our preferences determine what is good or bad for us: is it our preferences at the time when we choose or at the time when the effects of our choices set in?

Sunstein recognizes that there are problems with this kind of conception of the good, but 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, p. 46). He does, however, then primarily seem to understand the relevant problems as *epistemic*, that it is often difficult to know what our ultimate preferences or judgments are, whereas the argument here is about the *constitutive* role of such preferences or judgments: the notion of ultimate preferences or judgments is not just somewhat impractical, it is misguided. We should reject the metaphysical notion of a definitive set of ultimate preferences that constitute what really is good for a person. But this also means that we cannot reasonably use a remnant of the old standard picture in determining which kinds of choices might be appropriate targets of paternalist interventions; we should not both reject that picture as a description and still use it as a yardstick. Instead, we need to develop a more complex picture of decision-making competence, one built on the realities of human psychology.

Human Development and Competence

While one can certainly see something like the three-component picture discussed above lurking in the background (and sometimes even being made explicit in the foreground) in much of philosophy and economics, the authors who make use of this kind of model have, of course, not believed that adult human beings are rational decision-makers through and through. And it is certainly true that in doing moral and political theory it would be unworkable to consider human decision-making in full empirical detail. We need to rely on idealizations. A good idealization should, however, arguably capture something like a core of the thing for which it is an idealization.[10] As noted above, even authors who emphasize how our choices are often determined by contextual factors in the end often operate with a model of human agency where there is something like an inner rational agent trapped in an

[10] Weisberg (2007) has distinguished between *Galilean idealization*, which is done for computational purposes and which depends on our current computational capabilities, and *minimalist idealization*, which strips away irrelevant details in order to facilitate understanding of a phenomenon. With respect to Galilean idealization the above point need not hold, misrepresentation is then not a problem as such: if the idealization in question enables us to, say, make better predictions than other models would, it is a reasonable idealization. This could potentially be the case in economics (although some behavioral economists might disagree), but idealizations in moral and political theory that play a constitutive role in determining what is rational or what is good, rather than predicting behavior, cannot reasonably be considered Galilean (since there is no independent yardstick by which to measure the computational success of such idealizations—they themselves are supposed to serve as yardsticks).

outer psychological shell (Sugden 2015). The kind of picture that seems to be at work is one of a continuum that runs from children to idealized constructs such as *Homo economicus*, and where adult human beings have developed enough to be sufficiently similar to *Homo economicus* to be treated, on the whole, *as if* they were such decision-makers. This is a relatively one-dimensional and linear view of how we develop as decision-makers.[11] It is a simple picture, but also a deeply problematic one. It is a picture where even though System 1 thinking dominates in everyday life, it is still conceptualized almost like an aberration.

While it might very well be the case that many adults are more adept at using System 2 reasoning than most children, the picture of decision-making competence that emerges from the empirical research should lead us to question the idea of a single unified continuum of human development. While many of the empirical results discussed above point to how we are prone to failures of rationality, many of these results arise when we are faced with kinds of choices that we do not make on an everyday basis. In fact, there is considerable evidence that for many everyday tasks, we actually succeed worse when we use System 2 thinking rather than System 1 thinking (Gigerenzer 2000, 2008). We are essentially competent in much of everyday life. It is just that the competences that we have largely consist in our System 1 responses having gradually become better calibrated through repeated performances. Rational thinking (in the traditional sense) might certainly at times be of help in such processes of calibration and recalibration, but its role will tend to be marginal compared to feedback from our social environment.

Even on the standard picture, the exact legal line that we draw for when we count people as adults will always be arbitrary, something clearly recognized by Feinberg (1986, p. 326): 'Such boundary lines as the eighteenth or twenty-first birthday are simply approximations (plausible guesses) for the point where *all* the person's decision-making capacities are fully matured.' This kind of remark does, however, still presuppose that there is a point where our development is completed: we have become rational decision-makers. On the alternative picture, however, if coming of age really were to depend on our rational faculties approximating those of *Homo economicus*, then none of us would be eligible. But this does not mean that we are not much more competent decision-makers at the age of 18 than at the age of 12, it is just that the difference instead primarily lies in having System 1 responses that are well-calibrated enough to be let loose in society.[12] If we are to frame it in terms of rationality, we could say that by then we have achieved a sufficient level of, to use a phrase employed by Gerd Gigerenzer, *ecological rationality*.[13] Such calibrated

[11] The six-stage account of moral development put forward by Kohlberg (1973) is another example of an account where a certain type of principle-driven rational agency is seen as the end-point of our development from children to adults.

[12] Whitehead (1911, p. 46) made a similar point on the societal level: 'Civilization advances by extending the number of important operations which we can perform without thinking about them. Operations of thought are like cavalry charges in a battle—they are strictly limited in number, they require fresh horses, and must only be made at decisive moments.'

[13] On this kind of picture, adaptive behavior will be about achieving a match between mind and environment: when behavior is adaptive, interactions with and within the environment will run smoothly. It is like two people dancing, we can see when they are in synch without positing any specific and

System 1 responses will often function well even when people are in many ways relatively uninformed and unreflective.

On the picture that emerges here, our decision-making competence, irrespective of whether we are children or adults, is patchy. The exact layout of this quilt of competences will depend on our concrete experiences and which kinds of choices we actually tend to make on a regular basis in everyday life. In addition to this, we are still certainly capable of occasional uses of System 2 thinking. The difference between the two pictures lies in whether we think of people as being primarily guided by immediate responses, with occasional uses of rational thought, or primarily guided by rational thought, but with occasional lapses due to gut reactions. The suggestion here is that we should go for the former picture. Such an approach still leaves room for questions about what is rational in the more traditional sense, in contrast to responses being ecologically rational, but part of what is rejected is the idea of some kind of true preferences or latent (fully informed and rationally formed) preferences, somehow hidden beneath the quilt of System 1 responses. This also means that System 1 responses cannot in the end be evaluated in the light of such true preferences, so this is not a two-level approach where ecological rationality is always ultimately to be cashed out in terms of responses being instrumentally rational in the long run. But neither is it an approach according to which System 2 reasoning is never applicable—it is just that it is merely occasionally applicable. More precisely, it is applicable when we can identify clear targets for optimization and have sufficient knowledge of how to achieve such optimal results. It is not just our exact preferences that are context-dependent, but also the relevance and applicability of System 2 thinking.

Given the above picture of two kinds of decision-making competence that are both limited in their applicability, we can safely assume that human life will not be so neatly ordered that every choice will be such that we are competent in handling it in one of these ways. Some choices will fall between the cracks and these choices then seem reasonable as candidates for someone else being able to choose more wisely for us. We can, however, not just assume that this automatically means that paternalist interventions will be justified. To begin with, the question of who is best suited to make a decision is always a question of relative levels of competence. This point was made by Mill, but Hayek (1937, 1945) put forward a more elaborate version of it in his classic argument against central planning and for free market solutions. He starts by acknowledging that as individuals we only have partial knowledge of our circumstances and only fairly limited rationality. It is just that the policymakers and bureaucrats will still be in an inferior position. They will ultimately rely on statistical knowledge, which even if accurate, does not include knowledge of the particular circumstances of individuals. This means that they still cannot surpass the effectiveness of the incremental modifications of behavior that individuals are capable of when reacting to changes in their circumstances. Another type of argument is that even if people are prone to failures, we are on a policy level

Footnote 13 continued
determinate end that they are trying to achieve through dancing. An early proponent of this kind of picture was Herbert Simon, see for example his (1990).

looking not at individuals, but at overall patterns. As Eugene Fama has put it with respect to anomalies that are not predicted by the efficient-market theory of financial trading: 'If anomalies split randomly between underreaction and overreaction, they are consistent with market efficiency' (1998, p. 284).[14] Both Hayek and Fama are strongly focused on the information-processing capabilities of individuals and markets and their arguments point to how the bare fact that we as individuals often fall short of ideals of rationality does not automatically justify interventions into our behavior.

We must, however, distinguish between two ways in which we can fall short of ideals of knowledge and rationality systematically and not just occasionally. First, we can have widespread but *dispersed* failures, i.e. we regularly fail to live up to the ideals, but we do so in a very wide range of ways and for lots of different reasons. If this is our situation one might perhaps expect, or at least hope, that our failures will even out in the long run, at least on the level of populations but perhaps even for many individuals. And even if they do not even out, the kind of general information that would-be interveners have to base their decisions on is at any rate unlikely to enable them to make better choices than people can themselves. Yet if our errors are not more or less randomly dispersed, but instead due to certain specific blind spots, we might instead have strongly *clustered* failures, i.e. we will have particular trouble spots where we systematically and predictably behave incompetently and where things will not even out because we tend to fail in the same way again and again. Given this latter type of scenario, if we can identify these problem areas, then even if we merely have general knowledge, it is still quite possible that we can regularly make better choices for people than they would themselves, at least on average. The Mill–Hayek line of argument provides us with an important reason to be wary of government overreach, but it is an argument that is formulated in a completely general way and we should not think of decision-making competence in that way. The argument here is instead that a well-functioning system of making and implementing public policy can have a relative advantage compared to people themselves, but also that whether this is actually the case will always have to be established through an area-specific analysis.

It should be said that the question here is not just whether there are such areas where paternalist interventions can be reasonable, but also about where the burden of proof lies. As already pointed out, having and exercising substantial self-rule over one's life can be understood as valuable both instrumentally and inherently, and self-rule in any substantial sense would be severely undermined if we did not on the whole have the freedom to make stupid choices. It accordingly seems reasonable that government should largely stay out of everyday life and that government has to earn its right to intervene with respect to us, rather than that we have to earn the right to be left alone. Not intervening should be the default. There are certainly those who take an even stronger position, namely that at least under normal

[14] Fama is optimistic about this actually being the case; for a more pessimistic take, see Shiller (2003). It should be pointed out that the financial markets constitute a special case; these are decision environments that involve highly quantifiable goods, highly knowledgeable agents, and high volumes of transactions (the latter is important because we communicate through our transactions). Things might accordingly even out in the financial markets without doing so in other decision-making environments.

circumstances our right to self-rule precludes all paternalist interventions,[15] but it is difficult to find anyone who thinks that even children should be accorded such an absolute right, which indicates that even on such a view, competence presumably must matter. And if competence varies, then why should not the value of being the one who effectively makes the decisions vary too? Even as adults there will always be areas where we have little or no experience as well as areas where our heuristics and biases will systematically tend to lead us the wrong way. In short: there will be areas where our competence is similar to the competence of a child. And since we usually find it fairly unproblematic to intervene into the lives of children, should we not have a similar attitude to the parts of our own lives where our System 1 competence is more akin to that of children?

Here it might be objected that even if certain interventions might, due to the uneven competency profile which people tend to have, make more sense than others, paternalistic interventions could still be incompatible with the kind of respect that we are owed as citizens in liberal societies.[16] As citizens we are located in ongoing moral and political discourses which involve treating each other as adults to whom things are not merely done, but who are owed justification; would it then not be odd if this kind of fundamental respect had an on–off character? There are, however, two ways in which children tend to be treated as children: (1) that decisions are taken for them, and (2) that they are often not given any real explanations or justifications for those decisions—many things just happen to them. Arguably, the respect we are owed as citizens in a liberal society is primarily about having an absolute right to justification rather than an absolute right to non-interference, i.e. as citizens we are always owed proper justifications for government policies.[17] But such a right is perfectly compatible with the present line of argument.

Identifying Areas for Legitimate Interventions

The idea here is that paternalist interventions can be legitimate policy options, but that the burden of proof falls on the one proposing such interventions. Since the kind of argument that then needs to be made is inevitably comparative, there is a need to identify principles or criteria based on which we make such comparisons. Given the argument above, it seems reasonable to say that what is needed is: (1) that we are dealing with an area where there are clear problems with our System 1 thinking as a way of dealing with what is at issue in that area, and (2) that there is a live

[15] Mill (1859) is the classic example and Feinberg (1989) a more developed version; as already discussed in the paper, both of them work with a different account of decision-making competence than the one proposed here, and where the difference between adults and children will, except for a transitionary gray area, be more thoroughgoing.

[16] Quong (2010, pp. 100–103) has provided this kind of Rawlsian argument against paternalist interventions, although it should be noted that his point is just that such policies are presumptively wrong.

[17] Exactly what counts as a proper justification is a further question, but at the very least it would not seem to simply be a question of whether the person accepts the reasons given—for instance, some people reject the legitimacy of the taxation necessary for a welfare state, but such taxation can arguably still count as justified (and people are not disrespected as citizens because they are taxed to the needed extent).

possibility of applying System 2 thinking on the level of policy-making with the aim of effectively making the relevant choices for us in a way that can make us better off. Each of these two points can in turn be broken down into two sub-points and the suggestion here is that all four need to be clearly satisfied if paternalist interventions are to possess a basic level of legitimacy as policy options.

To begin with we should, as already alluded to, be dealing with areas where there is *clustering of problematic tendencies* in System 1 decision-making (as well as it being unlikely that we will make up for this problem by use of System 2 reasoning). For instance, given our optimism and overconfidence and our poor skills at handling probabilities and large numbers, we are especially likely to make mistakes in areas that involve long-term cumulative effects and where the link between particular decisions and the overall effect is difficult to see, even though on a statistical level the link is clearly discernible.[18] These are areas of choice where the grounds for the normal relative advantage of individuals compared to policy-makers and bureaucrats will no longer obtain: an opening for compensatory paternalist measures is accordingly created. While the cases that will be most reasonable for intervention are ones where several problematic System 1 tendencies are all potentially present and where there is accordingly a high risk of irrational or incompetent behavior, this is not to say that everyone will necessarily fail at securing what's best for them, but simply that there are clearly discernable risks on a population level.

Many health-affecting habits, e.g. eating junk food, drinking heavily, smoking, not exercising, are characterized by cumulative effects where individuals might be prone to wishful thinking or weakness of will, but where on a general statistical level, the correlations are fairly straightforward. It should be clear that a lot of these habits involve a high discrepancy between System 1 and 2 responses; in fact, the individual herself might during calm reflection very well come to the conclusion that she should break her habits, but when System 1 kicks in again, the habits remain. Many of the relevant effects can also involve a form of intrapersonal tragedy of the commons: one cigarette, one drink, or one supersized fast-food meal will not make any difference, but in the long run, certain life-patterns will have detrimental effects. Government regulations can at least potentially play a helpful role in such cases.

Secondly, we should be dealing with behavior that takes place in *low-learning environments*. We all make mistakes and not only is there no way of stopping us from making any mistakes, it would not even be desirable to fully do so. Even if we are dealing with areas where many problematic tendencies are involved, being allowed to make mistakes can still be an important part of our processes of learning and growth. The experience of sometimes failing can also be a valuable component in forming realistic expectations about what you can get out of life. However, sometimes we make mistakes that we never reliably learn are mistakes even if they are, or that we will not learn about until it is too late to learn from them. Areas of

[18] Le Grand and New (2015, pp. 82–101) distinguish between four types of reasoning failure (limited technical ability, limited imagination/experience, limited willpower, and limited objectivity), but their model is formulated in terms of how we fall short of the standard account of rational choice and, as argued here, we should not assume that rational choice empties out what decision-making competence can be about.

choice where there is a significant distance in time or space between our actions and their full effects are especially prone to be low-learning environments since the feedback mechanisms will tend to be weak. The number of relevant interactions that we have is also very important; good gut reactions are formed gradually and through solid experience.[19] In a high-learning environment, we can expect System 1 thinking to gradually adapt so that it becomes a reliable way of making reasonably good decisions,[20] whereas in a low-learning environment this will not happen to the same extent. Additionally, if we are in a high-learning environment and still do not adapt there is a stronger case to be made for our failures simply being our own responsibility. The case for government intervention accordingly becomes weaker.

It should be said that even in what is generally a high-learning environment, there might be certain decisions which we have few opportunities to make and, then, that environment will be a low-learning one with respect to them, e.g. even a person who has spent considerable time behind the wheel and has excellent immediate responses in many traffic situations might still not be able to appreciate the effects that not having a seat-belt will have once you crash your car. Mandatory use of seat-belts in cars and helmets for motorcyclists are examples of paternalist interventions that are more or less universally accepted in contemporary industrialized countries. Our economic life contains similar safety measures, such as pension plans and insurance schemes, which seem like clear analogues to seat-belts. Maybe the crashes will not be quite that literal, but there is certainly a risk of crashing. The relevant learning environments will often be quite poor in relation to such economic safety measures as well.[21] There is no room here for a full analysis of any specific paternalist measures, but at the very least a safety measure such as mandated seat-belts would seem to clear all four of the criteria proposed here.

Thirdly, on the government side of things we should be dealing with areas where *policy-making can ultimately be evidence-based*, i.e. areas where we can assess the impact of different policy measures on the policy targets. This is not the same thing as demanding that we must know in advance that a given policy measure will work well. Even if it is reasonable that there is a general presumption against intervening, to demand certainty or near-certainty would be to go beyond merely accepting that kind of presumption. There is a balance to be struck here. It is reasonable to insist on evidence-basing, but this demand should not be misused in order to slip a block of more or less all interventions in through the back door. Not only do we never get absolute certainty in politics, but inaction is always also a choice that comes with its own risks. The choice is rarely between a non-intervention that we know will work well and an intervention that might work well. The rising prevalence of obesity in

[19] Thaler and Sunstein identify low frequency of choices and poor feedback as two problems (2008, pp. 128–130), but it is arguably the overall low-learning character of an area of choice that is the core issue for the question of legitimacy—then there might be a variety of reasons as to why something is a low-learning environment.

[20] For a discussion of what characterizes high-learning environments, see Kahneman and Klein (2009).

[21] It should be acknowledged that there are other factors at play here. For instance, when it comes to systems where it is optional to purchase health insurances, insurability and affordability are important factors as well. But they cannot fully explain why some people remain uninsured (Bundorf and Pauly 2006), even though it seems quite obvious that it is a risk not worth taking.

many countries (OECD 2017) is an example of this. There might not be any specific potential measure which, based in our present knowledge, are guaranteed to curb this development, but current trends point to a situation where the economic and societal costs are likely to be significant if this development is not curbed. Here we know that inaction will come with certain serious consequences.

In order to build knowledge about which policies that do work, policy-making must at times be allowed to be experimental. We occasionally need to try things out for the first time. It is then important, in order for experimental policy-making to still be responsible policy-making, to at least have a *research-based* idea about the mechanisms through which the intended policy is to achieve its goals as well as an idea about how to measure the planned policy's effects. Interventions will often tend to target factors that increase the risk of some bad or problematic effect. For instance, it appears fairly well-established that there is a link between, on the one side, high sugar intake and, on the other, obesity and type 2 diabetes (Hu 2013). It is not difficult to come up with policy measures that target the sugar-intake part of this mechanism, but however confident we are in this link being not just a correlation but also a causal connection, there is still a question about which such measures that will actually have an effect on the total sugar intake (rather than just shift consumption between different sources). This particular area is one where there is already experimental policy-making going on and we are gradually gaining knowledge about what works,[22] but the general argument here is simply that policy-making can be governed by an ideal of evidence-basing and still be experimental up to a point.

Finally, we should be dealing with areas where the effects that policies can have on individuals are such that they have a clear connection to *important and relatively noncontroversial values* for the individuals affected by the policy. What this means is that it is not enough that there are reliable indicators by which the success of the policy can be measured in terms of effectiveness and efficiency and with which we are able to follow changes over time in the population. Measurable effects must also translate to real gains. To be able to assess this we need an account of well-being or the human good and, as already argued, the type of account that compensatory paternalists tend to favor, namely some form of preferentialism, has turned out to be deeply problematic. Apart from it not being realistic to have policies that target the specific preferences of each individual, the very idea that there are definitive answers, even if only in principle, to what each person really wants is misguided. There are, of course, other long-standing conceptions of well-being as well, such as hedonism or some form of objective-list account,[23] but one problem with many such accounts is that they would themselves involve an element of paternalism: telling individuals what really matters to them. There are, however, also other options, accounts of the good that leave more room for individuals to form their own conceptions of the good and act on them—*primary goods* (Rawls 1999) and

[22] For instance, see Colchero et al. (2016) for indications that taxing sugar-sweetened beverages has some effect on behavior when it comes to sugar intake.

[23] Objective-list accounts can take many forms; for a leading example see Griffin (1986, p. 67).

capabilities (e.g. Nussbaum 2001) being two main options.[24] It should be noted that with respect to the argument here, the choice of a specific account of the good is about choosing a plug-into the overall framework, which can work with primary goods just as well as with capabilities. But we need *an* account of the good for there to be translatability between measurable policy outcomes and something that counts as a real benefit to the people affected.

Having said this, the framework of primary goods does arguably have certain advantages given the conception of human decision-making advocated here. While Rawls's own way of setting up his list of primary goods is certainly problematic since it is governed by ideas such as supposing that 'each person has a rational plan of life drawn up subject to the conditions that confront him' (1999, p. 80), what arguably does most of the work in identifying which things that are primary goods is not so much the planning part (which is psychologically quite unrealistic) but the variability of what different people seek and the versatility and pronounced usefulness of certain goods. As long as we have a conception of human decision-making that recognizes that we, for the most part, are competent in what we seek and what we do (and such an account has been proposed here), then even if there might be significant indeterminacies with respect to specific preferences, we should still be able to make certain generalizations, based in actual behavior, about which things are to be regarded as primary goods under normal circumstances. Additionally, identifying something as a primary good does not involve taking a stand on whether it is just an instrumental good or also a final one; what we are taking a stand on is simply whether it is a reasonable goal of public policy (which is really a stance we need to take anyway, at least if we favor there being some scheme of distributive justice in place). With respect to paternalist policy-making, things such as wealth and income or health[25] are fairly straightforward to track statistically, but are also things that have the relevant kind of versatility and generic importance that make them reasonable to understand as primary goods.

Some Final Remarks

If all four of the above criteria are clearly satisfied, the idea here is that there is a basic moral and political legitimacy of paternalist interventions. This does, however, not settle the question of whether a specific intervention should at the end of the day actually be put into practice. It will still have to be assessed in terms of the costs and benefits of that particular policy (Conly 2013, pp. 150–152). In doing this kind of analysis the key balance to strike will be about how much benefit in terms of well-being can be had and at what cost in terms of limitations to our self-rule. There will hardly be any easy way of measuring and comparing different

[24] For a number of good essays that provide development of, and comparisons between, these two options, see Brighouse and Robeyns (2010).

[25] Health is a more complex matter than wealth/income, since it is clearly a multi-dimensional goal and it is not obvious how it is to be broken down into components and measured, or how different aspects of health are to be comparatively valued. At the same time, many of these problems will mainly lie on the level of principle and should often be reconcilable on a more practical level (Wolff et al. 2012).

interventions in terms of how much they circumscribe our self-rule, but in general it seems reasonable to say that nudges tend to be less invasive than coercive measures and if we recognize a general right to non-interference, it could very well make sense to opt for a less invasive measure even if it also comes with less benefit—at least as long as the difference is not significant. It should also be kept in mind that the overall number of interventions must be considered as well. The above model is a model for assessing specific interventions, so by itself it does not safeguard against the risk that our self-rule falls prey to death by a thousand cuts. The idea here has been that being treated as adults *on the whole* is compatible with there being some choices that are effectively taken away from us, but even small steps can in principle take us across a line where it is no longer reasonable to say that people are being treated as adults on the whole. So even if the above model works as a way of determining the basic legitimacy of paternalist policy options, paternalist policy-makers could still be faced with a choice of which interventions are the most important to make and settle simply for these.

A further complicating factor that should be kept in mind is that on the conception of human agency and decision-making competence outlined above, human beings systematically tend to overestimate their competence. This means that there will be a general tendency that perceived legitimacy might not match up with legitimacy according to these four criteria. At the very least, people can probably be expected to think that while other persons need paternalist interventions into their lives, *I* am most certainly capable of deciding for myself. Effective policy-making often relies on at least some perceived legitimacy among those affected by the policy, since otherwise active resistance to the policy might become widespread; so in general it will be wise to focus on interventions for which a reasonable level of such perceived legitimacy can be built and this can be one important factor when it comes to choosing the right type of paternalist intervention in an area where the use of some kind of such intervention is prima facie legitimate. It should, however, be remembered that there are always pedagogical concerns when it comes to enacting different policies, so paternalist interventions do not introduce a new dimension into politics. But their controversial character certainly emphasizes why it is reasonable to hold that the burden of justification lies on government, and that it needs to be shown that paternalist interventions take place in areas where there is both a trouble spot for individual human decision-making and, so to speak, a sweet spot for reasoned policy-making.

Acknowledgements The author completed this article while holding a position as Research Fellow granted by the Royal Swedish Academy of Letters, History and Antiquities, as well as greatly benefiting from participation in the meetings of the Bank of Sweden Tercentenary Foundation Program on Science and Proven Experience. An earlier version of the text was presented at the Research Seminar in Practical Philosophy and Political Theory, Gothenburg University. Many thanks to the participants there, as well as to the two referees for *Res Publica*, for their helpful comments.

References

Anderson, Joel. 2014. Regimes of autonomy. *Ethical Theory and Moral Practice* 17: 355–368.

Aristotle. 1999. *Nicomachean ethics*. trans. T. H. Irwin. Indianapolis, IN: Hackett.

Benhabib, Jess, Alberto Bisina, and Andrew Schottera. 2010. Present-bias, quasi-hyperbolic discounting, and fixed costs. *Games and Economic Behavior* 69: 205–223.

Brighouse, Harry, and Ingrid Robeyns (eds.). 2010. *Measuring justice: Primary goods and capabilities*. Cambridge: Cambridge University Press.

Bundorf, Kate, and Mark Pauly. 2006. Is health insurance affordable for the uninsured? *Journal of Health Economics* 25: 650–673.

Colchero, M.Arantxa, et al. 2016. Beverage purchases from stores in Mexico under the excise tax on sugar sweetened beverages: Observational study. *BMJ* 352: h6704.

Conly, Sarah. 2013. *Against autonomy*. Cambridge: Cambridge University Press.

Fama, Eugene. 1998. Market efficiency, long-term returns, and behavioral finance. *Journal of Financial Economics* 49: 283–306.

Feinberg, Joel. 1986. *Harm to self*. New York, NY: Oxford University Press.

Fetherstonhaugh, David, et al. 1997. Insensitivity to the value of human life: A study of psychophysical numbing. *Journal of Risk and Uncertainty* 14: 283–300.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue. 2002. Time discounting and time preference: A critical review. *Journal of Economic Literature* 40: 351–401.

Gigerenzer, Gerd. 2000. *Adaptive thinking: Rationality in the real world*. New York, NY: Oxford University Press.

Gigerenzer, Gerd. 2008. *How people cope with uncertainty*. New York, NY: Oxford University Press.

Grether, David, and Charles Plott. 1979. Economic theory of choice and the preference reversal phenomenon. *American Economic Review* 69: 623–638.

Griffin, James. 1986. *Well-being*. Oxford: Clarendon Press.

Haidt, Jonathan. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108: 814–834.

Hall, Lars, Petter Johansson, and Thomas Strandberg. 2012. Lifting the veil of morality: Choice blindness and moral attitude reversals on a self-transforming survey. *PLoS ONE* 7: e45457.

Hayek, Friedrich. 1937. Economics and knowledge. *Economica, New Series* 4: 33–54.

Hayek, Friedrich. 1945. The use of knowledge in society. *The American Economic Review* 35: 519–530.

Hogarth, Robin, and Hillel Einhorn. 1992. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology* 24: 1–55.

Hu, Frank. 2013. Resolved: There is sufficient scientific evidence that decreasing sugar-sweetened beverage consumption will reduce the prevalence of obesity and obesity-related diseases. *Obesity Reviews* 14: 606–619.

Kahneman, Daniel, and Gary Klein. 2009. Conditions for intuitive expertise: A failure to disagree. *American Psychologist* 64: 515–526.

Kohlberg, Lawrence. 1973. The claim to moral adequacy of a highest stage of moral judgment. *The Journal of Philosophy* 70: 630–646.

Le Grand, Julian, and Bill New. 2015. *Government paternalism*. Princeton, NJ: Princeton University Press.

Longoa, Matthew, and Stella Lourenco. 2007. Spatial attention and the mental number line: Evidence for characteristic biases and compression. *Neuropsychologia* 45: 1400–1407.

Merten, Katharina, and Andreas Nieder. 2009. Compressed scaling of abstract numerosity representations in adult humans and monkeys. *Journal of Cognitive Neuroscience* 21: 333–346.

Mill, John Stuart. 1989 [1859]. *On liberty*. In *On liberty and other writings*, ed. Stefan Collini. Cambridge: Cambridge University Press.

Montier, James. 2007. Behaving badly. In *Behavioural investing: A practitioner's guide to applying behavioural finance*. Oxford: Wiley.

Nussbaum, Martha. 2001. *Women and human development*. Cambridge: Cambridge University Press.

OECD. 2017. *Obesity update*. https://www.oecd.org/els/health-systems/Obesity-Update-2017.pdf. Accessed 16 Jan 2018.

Quong, Jonathan. 2010. *Liberalism without perfection*. Oxford: Oxford University Press.

Rawls, John. 1999. *A theory of justice*, rev ed. Cambridge, MA: Harvard University Press.

Shiller, Robert. 2003. From efficient markets theory to behavioral finance. *The Journal of Economic Perspectives* 17: 83–104.
Simon, Herbert. 1990. Invariants of human behavior. *Annual Review of Psychology* 41: 1–19.
Slovic, Paul. 1995. The construction of preference. *American Psychologist* 50: 364–371.
Slovic, Paul. 2007. 'If I look at the mass I will never act': Psychic numbing and genocide. *Judgment and Decision Making* 2: 79–95.
Strotz, Robert. 1955–56. Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies* 23:165–180.
Sugden, Robert. 2008. Why incoherent preferences do not justify paternalism. *Constitutional Political Economy* 19: 226–248.
Sugden, Robert. 2015. Looking for a psychology for the inner rational agent. *Social Theory and Practice* 41: 579–598.
Sunstein, Cass. 2016. *The ethics of influence*. Cambridge: Cambridge University Press.
Svenson, Ola. 1981. Are we less risky and more skillful than our fellow drivers? *Acta Psychologica* 47: 143–151.
Thaler, Richard. 1981. Some empirical evidence on dynamic inconsistency. *Economics Letters* 8: 201–207.
Thaler, Richard, and Cass Sunstein. 2003. Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review* 70: 1159–1202.
Thaler, Richard, and Cass Sunstein. 2008. *Nudge*. Yale, CT: Yale University Press.
Tversky, Amos, and Daniel Kahneman. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263–292.
Tversky, Amos, and Itamar Simonson. 1993. Context-dependent preferences. *Management Science* 39: 1179–1189.
Tännsjö, Torbjörn. 2015. Context-dependent preferences and the right to forgo life-saving treatments. *Social Theory and Practice* 41: 716–733.
Wason, Peter. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology* 12: 129–140.
White, Mark. 2013. *The manipulation of choice*. New York, NY: Palgrave Macmillan.
Whitehead, Alfred. 1911. *An introduction to mathematics*. London: Williams & Northgate.
Weinstein, Neil. 1980. Unrealistic optimism about future life events. *Journal of Personality and Social Psychology* 39: 806–820.
Weisberg, Michael. 2007. Three kinds of idealization. *The Journal of Philosophy* 104: 639–659.
Wolff, Jonathan, et al. 2012. Evaluating interventions in health: A reconciliatory approach. *Bioethics* 26: 455–463.