View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Warwick Research Archives Portal Repository

Original citation:
Mills, Chris. (2015) The heteronomy of choice architecture. Review of Philosophy and Psychology, 6 (3). pp. 495-509.

Permanent WRAP URL:
http://wrap.warwick.ac.uk/91069

Copyright and reuse:
The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:
"The final publication is available at Springer via http://dx.doi.org/10.1007/s13164-015-0242-7

A note on versions:
The version presented here may differ from the published version or, version of record, if you wish to cite this item you are advised to consult the publisher's version. Please see the 'permanent WRAP url' above for details on accessing the published version and note that access may require a subscription.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications

The Heteronomy of Choice Architecture

Review of Philosophy and Psychology 6, no. 3 (2015): 495-509.

Abstract

Choice architecture is heralded as a policy approach that does not coercively reduce freedom of choice. Still we might worry that this approach fails to respect individual choice because it subversively manipulates individuals, thus contravening their personal autonomy. In this article I address two arguments to this effect. First, I deny that choice architecture is necessarily heteronomous. I explain the reasons we have for avoiding heteronomous policy-making and offer a set of four conditions for non-heteronomy. I then provide examples of nudges that meet these conditions. I argue that these policies are capable of respecting and promoting personal autonomy, and show this claim to be true across contrasting conceptions of autonomy. Second, I deny that choice architecture is disrespectful because it is epistemically paternalistic. This critique appears to loom large even against non-heteronomous nudges. However, I argue that while some of these policies may exhibit epistemically paternalistic tendencies, these tendencies do not necessarily undermine personal autonomy. Thus, if we are to find such policies objectionable, we cannot do so on the grounds of respect for autonomy.

Keywords: Heteronomy, Choice Architecture, Personal Autonomy, Epistemic Paternalism.

1. Introduction

Choice architecture (or nudging) is an approach to policy design that seeks to harness evidence from behavioural economics and cognitive psychology to overcome blunders we commonly make in our decision-making. These include a reliance on heuristics (such as anchoring, availability, and representativeness) and biases toward unrealistic optimism, preservation of the status quo, loss aversion, and vulnerability to framing effects (Thaler & Sunstein 2008, pp. 24-

40).[1] Sometimes these phenomena are harmless and we employ them as shortcuts in our everyday practical reasoning without problem. However, sometimes these phenomena lead to suboptimal outcomes, not only in terms of some objective measure of value, but even according to our own subjective standards of success. We act as boundedly-rational agents, sometimes to our detriment.

Once identified, policy-makers face a choice about how to design policies in light of these cognitive shortcomings. A nudge is a particular type of policy which seeks to bring about beneficial ends by either exploiting or preventing these biases. Some nudges are paternalistic because they are intended to promote the well-being of the subject because he or she is judged incapable of doing so themselves.[2] These nudges, offered under the banner of libertarian paternalism, are a specifically motivated subset of choice architecture. These policies face questions about the legitimacy of both their means and their ends.

However not all nudges are paternalistic. Some are intended to overcome collective action problems, prevent large-scale harms or bring about socially just outcomes. For example, nudges may be designed to increase organ-donation, combat climate change, or reduce discrimination in the work place. As these policies seek to aid third-parties (even after the death of the subject in some instances) they are cases of non-paternalistic nudging. Here I am interested in this latter group of policies. By engaging with these cases I seek to avoid recycling many of the traditional concerns about paternalism. Instead I am assuming the legitimacy of the government's aims and questioning whether the use of choice architecture as a means to securing these legitimate aims is necessarily disrespectful toward personal autonomy.

[1] In his most recent work on choice architecture, Sunstein (2014b, pp. 34-50) describes these phenomena as 'behavioural market failures'. He groups these into four distinct sets: 1) present bias, time inconsistency and inter-temporal internalities; 2) saliency and shrouded attributes; 3) unrealistic optimism; and 4) problems with probability and availability.

[2] Here I characterise paternalism as a motivational wrong. This characterisation is contestable. For more on the plausibility of this characterisation and paternalism's justificatory burden, see Mills 2013b.

In response to this question I seek to develop and defend a prior claim that some instances of choice architecture are not only compatible with personal autonomy, but can promote it (Mills 2013a). To defend this claim I begin by outlining our reasons for avoiding heteronomous policies (§2). I then discuss four features of nudging that are salient to autonomy-based concerns and explain why they should lead us to believe that some instances of choice architecture are non-heteronomous (§3). I show this claim to be compatible with contrasting conceptions of personal autonomy and, further, how it can be extended to the promotion of autonomy (§4). I then propose a related epistemic criticism of nudging that threatens to undermine the claim that nudging respects autonomous choice (§5). I defuse this argument before briefly concluding (§6).

2. Choice Architecture and Personal Autonomy

The normative standard most commonly employed to assess the permissibility of choice architecture is the preservation of freedom of choice. Libertarian proponents of choice architecture emphasise that nudges preserve freedom of choice because individual choices are not coercively restricted. This claim has been a major selling-point of the approach and is responsible for much of its initial popularity with policy-makers. However this standard of permissibility faces a number of concerns. First, whether relatively opaque non-coercive influences are all things considered preferable to more transparent but more coercive efforts is unclear (e.g. Conly 2013, pp. 29-36). Further, the truth of the non-coerciveness claim hinges on the libertarian tendency to favour thin, negative conceptions of freedom. As a result, it may not hold across all conceptions of freedom. (Goodwin 2012, p. 88; Grüne-Yanoff 2012, p. 638).

My concern differs from this line of critique. Even if choice architecture does preserve liberty, this does not fully vindicate the method as respectful toward the choices of autonomous agents. We ought to concern ourselves with more than the number of options facing individuals. We should also consider the quality of those options and the individual's ability to reflect on these options in an authentic fashion. These autonomy-based concerns are at least equally

important as those concerning freedom of choice, but they have enjoyed far less detailed analysis.[3]

Personal autonomy is the capacity for an individual to determine and pursue her own conception of the good according to her own will.[4] There are a range of reasons why we might value this capacity instrumentally (e.g. we might think it good for a stable and progressive society). Further, we might value it intrinsically (e.g. as a necessary component of well-being or meaningful agency). Given this range of reasons, autonomy is often thought to be worthy of respect and to play a central role in our everyday moral conduct (e.g. Dillon 1995; Hill Jr. 2000; Kerstein 2013).

If we recognise the value of autonomy, reflecting on its content provides us with clear guidelines for respecting others. Although consensus on this reflection is not unanimous (more on this in §4), there is widespread agreement that the autonomous pursuit of the good requires a distinct combination of *internal* and *external* conditions. The internal conditions of personal autonomy primarily concern an individual's competency at decision-making and her independence from internal authenticity-threatening factors (e.g. phobias). The external conditions of personal autonomy primarily concern the quality of her options and her independence from external authenticity-threatening factors (e.g. coercion). Recognising these factors allows us to identify situations where respect for personal autonomy is at risk, and debate over how these factors are best fleshed out allows us to better employ respect for autonomy as a standard for permissibility.

Acknowledging the value of autonomy gives us reason to avoid heteronomy. Heteronomous behaviour can be caused by any reason for action that motivates

[3] This is partly due to Sunstein and Thaler's repeated refusal to engage at any great length with autonomy as an intrinsically valuable consideration (Sunstein & Thaler 2003, p. 1167, n. 22; Sunstein 2014b, p. 134). For the most detailed attempt, see Sunstein forthcoming.

[4] For more on the various other ways we might define personal autonomy (and whether it can be characterised in one single way), see Feinberg 1989.

an individual contrary to (e.g. by overriding or subverting) their authentic will.[5] Heteronomy specifically threatens the independence of an individual's will by disregarding her decision-making competency, thus bypassing part of what makes her decision her own. It is whether choice architecture shows this particularly disrespectful characteristic that concerns us. If it does, then even non-paternalistic nudges threaten a *pro-tanto* wrong against the subject. This would severely weaken the case for nudging.

Critics may suggest that choice architecture is necessarily heteronomous because it seeks to exploit heuristics and cognitive biases in our reasoning. Accordingly, choice architects pursue a programme of manipulation that undermines the independence of an autonomous agent's will by subverting the flaws in her decision-making competency to bring about particular outcomes. Even though nudges may leave the number of options in a subject's choice set unchanged, the choice architect exerts objectionable pressure on the individual's will to direct her behaviour within that choice-making scenario (e.g. Bovens 2009, p. 209; Hausman & Welch 2010, p. 28; Wilkinson 2013, p. 347; White 2013, p. 95). If the subject acts as the choice architect intends due to the pressure they experience (i.e. if the nudge succeeds), then the nudge undermines her autonomy by contravening the independence condition. In what follows, I contest two versions of this objection.

3. The Salient Characteristics of Choice Architecture

If the charge of heteronomy was true then choice architecture would be objectionable on grounds of respect for autonomy. The charge arises because the success of choice architecture relies on the very pressure that troubles critics. To

[5] On a strictly Kantian definition, heteronomous motivation is based on an impulse foreign from our reason, including inclinations (Kant 2012, 4:444). As such impulses do not originate from our rational will they prevent us from self-legislation. Further, because such impulses are not necessarily shared by all rational agents, they cannot justify universal categorical imperatives. This Kantian understanding of heteronomy is slightly narrower than the one that motivates my concern. Kantians interpret heteronomy through their interest in the possibility of moral autonomy and universal categorical moral obligations. I am merely concerned with sources of reasons for action that override the capacity for personal autonomy and I adopt the term with this broader usage in mind.

defuse this objection it must be shown that the pressure required for a nudge to be effective need not be heteronomous. To do this we must show that a policy can influence the subject's decision (e.g. counterfactually if the nudge were not in place the individual would have chosen differently) but that this influence does not entirely override her reasoning nor circumvent her decision-making competency. Previously, I have briefly sketched such an argument (Mills 2013a). If successful, it suggests that this criticism is too quick and that some instances of choice architecture can be both successful and non-heteronomous. Here I seek to develop and defend this argument, which revolves around four characteristics first suggested by Thaler and Sunstein as characteristics of good nudging. These characteristics, I suggest, are salient to respect for autonomy and should calm critics' worries.

The first characteristic is that choice architecture should be *primarily intended to facilitate an individual's pursuit of her own goals* (Thaler & Sunstein 2008, pp. 5-6).[6] This ensures that nudges must allow the subject significant freedom to select and pursue her own authentically adopted ends. This characteristic is crucial. It ensures that the subject's will becomes the lodestar for good nudge design, giving the policy-maker reason to design nudges that facilitate a subject's otherwise thwarted attempts at autonomous action. For example, such policies might reduce the subject's exposure to misinformation or offer helpful suggestions of ways of achieving their goals, thus aiding authentic behaviour. So long as the nudge tracks the subject's autonomous will (and does not contravene it) the nudge will avoid heteronomy.

Critics may object, however, to the difficulties of designing such policies. This condition requires the choice architect to have epistemic access to an individual's subjective standards. White suggests that the evidence required is unlikely to be available, thus leading the policy-maker into an objectionable process of value substitution (White 2013, pp. 64-79). Rebonato suggests the stronger claim that nudging in-line with an individual's own preferences is

[6] This ensures that when employed paternalistically, nudges are instances of *means* paternalism rather than *ends* paternalism (Sunstein 2014b, p. 19).

impossible for the very reason that allows nudging in the first place – the split nature of our reasoning (Rebonato 2012, pp. 153-158). The distinction between systems of reasoning that nudging takes for granted ensures that the architect cannot take an individual's decisions at face value. Some decisions will be blunders and others will not. Because of this, the validity of the subject's preferences is opaque to both the architect and the subject. How can we design policies that use the subject's tendency toward blunders to help her authentic decision-making when neither the architect nor the subject can be sure which decisions represent which?

The response to this problem has two parts. In the first instance, the philosopher can help. What distinguishes blunders from non-blunders is whether the decision furthers our subjective standard of success. That is, whether they are authentic to our conception of the good. So to design an effective and respectful nudge, we must identify cognitive factors that lead us to inauthentic blunders and seek to compensate for their presence. This requires our policy to reference a compelling account of authenticity e.g. one based on volitional necessities (Frankfurt 1982; 1998; Watson 2004) or coherence of preferences (Ekstrom 1993; 2005a; 2005b; 2010). To illustrate, we might design a choice prompt that asks choosers to consider what outcomes they feel that they could not live without or how their decision will compare to those that they usually make. These simple questions nudge the subject to consider her current decisions against her authentic motivations and preferences. With an account of authenticity in hand, the choice architect can come to understand which of the subject's decisions are blunders and which are not. If their policies enable (or do not prevent) authentic decisions, then their policies pass the test of effectiveness and respect.

Taking the second step of putting this into practice will likely require us to draw on the psychology of debiasing. Debiasing techniques are designed to help shrink the distance between systems of reasoning by drawing a subject's attention to present biases, rather than influencing her behaviour by replacing one bias with another (e.g. Jolls & Sunstein 2006; Pi et al. 2014). Such techniques are central to

reducing instances of unreflective blunders because they help the subject to identify and reflect on influences on their reasoning. When combined with an account of authenticity, debiasing techniques will allow the policy-maker to reduce decisions that the subject deems alienating, further aiding their autonomy (Trout 2005, p. 414).

The second salient characteristic is that choice scenarios should be designed to include *an acceptably low opt-out cost*.[7] This will allow the subject to avoid the policy if they feel threatened by its pressure or do not will the intended outcome. This characteristic requires the nudge to be both effective and easily avoided if the subject so desires, adding an additional test to contend with. Rebonato suggests that nudges cannot be both effective and avoidable (2013, pp. 200-209). This is because the effectiveness of choice architecture relies on a functional distinction between *nominal* and *real* freedom of choice. For a nudge to be successful, it must exert pressure in a way that reduces real freedom of choice, leaving only a nominal form of freedom of choice in its place:

> 'So, if the nudges of the libertarian paternalists – such as changing the default option – are effective, and exploit the decisional inertia of the choosers, then it makes little difference that there is a nominal right to opt out. And if their nudge is *very* effective, then having the nominal right to reverse the nudge makes *very* little difference'. (Rebonato 2013, p. 203, italics original).

Rebonato is correct to identify the possible tension between effectiveness and avoidability as relevant to the normative assessment of nudging. However, he mischaracterises this tension because of the ambiguity in how we might characterise the opt-out clause and its relationship with autonomy. To respect autonomy, nudges need to be *avoidable* rather than *reversible* (Sunstein 2014b, p. 60). Choice architects should ensure that there is a negligible cost to opting out (i.e. one that does not impair the voluntariness of an individual's actions) rather

[7] For an interesting argument concerning how the opt-out clause may combine with the intentions of the choice architect, see Wilkinson 2013, pp. 351-353.

than a non-negligible cost that can be compensated for or reversed at a later date. Reversibility allows a coerciveness in nudging (that directly contravenes the independence condition of autonomy) that avoidability does not.

With this clarification in mind we can see that the tension that Rebonato identifies does not prohibit all forms of choice architecture. A tension between effectiveness and avoidability exists when the aim of a nudge runs contrary to an individual's will. In this scenario the effective aim of the nudge differs from the subject's wishes, putting it in tension with the subject's desire to opt-out. If a choice architect attempts to alter individual behaviour in that manner, avoidability (via an opt-out) would be required to ensure respect for choice. But, as Rebonato suggests, the effectiveness of the nudge ensures that the individual cannot make use of the opt-out because the nudge undermines her reflective capacity. This is one way that heteronomous nudges fail to respect personal autonomy; either i) the policy succeeds and the individual wanted to opt-out but could not, or ii) the policy fails because the individual opts out.

However, as the previous characteristic suggests, not all nudges do this. Both the aim of the policy and how the policy achieves that aim matters here. Effectiveness is not determined by a single measure of pressure on the will but rather by a pluralistic measure of affective influence. Nudges affect our behaviour in more than one way. A distinction between types of influence can determine whether a policy is heteronomous or not (Blumenthal-Barby 2013, p. 192). As a result, there are various types of influence available to choice architects that allow nudges to be both effective and non-heteronomous. When a policy effectively constrains an individual's authentic pursuit of her goals, Rebonato's tension is pervasive but when such policies effectively facilitate her pursuit of her own goals, Rebonato's tension dissolves. As such, the opt-out condition gives the subject extra defence against poorly designed nudges by making it less likely that they will be effective. In contrast, well designed nudges (e.g. according to the previous characteristic) will not generate this tension.

The final constraints on the permissibility of nudging are provided by a pair of conditions - *publicity* and *transparency*. Thaler and Sunstein commit themselves to a loose Rawlsian principle of publicity (Thaler & Sunstein 2008, pp. 244-245). Rawls' publicity condition is part of his social contract approach to establishing principles of justice and, as such, is intended to represent part of the value of agreement over moral principles.[8] Although nudges do not require actual consent, publicity is relevant to ensuring respect for autonomy.[9] There are two ways of interpreting this requirement:

In its stronger form, *publicity* acts as the basis for Rawls' public justification requirement (Rawls 2001, pp. 26-29; Rawls 2005, p. 226; Thaler & Sunstein 2008, p. 245). This condition concerns the policy-maker's ability and willingness to justify her policies to those affected. It restricts the range of reasons that policy-makers can permissibly appeal to down to those that other individuals can reasonably be expected to endorse. If an instance of choice architecture is to meet this stronger constraint, the choice architect must have reasons for the intervention that they believe others are likely to share.[10]

In its weaker form, a commitment to publicity entails some form of *transparency*. Publicity requires the individuals affected to understand the policy as if they had agreed to it even if they haven't actually done so. This requires that the decisions

[8] In Rawls' original expression of his theory, he states that according to the publicity condition: 'The parties assume that they are choosing principles for a public conception of justice. They suppose that everyone will know about these principles all that he would know if their acceptance were the result of an agreement. Thus the general awareness of their universal acceptance should have desirable effects and support the stability of social cooperation.' (Rawls 1971, p. 133). Larmore suggests the following interpretation: 'The point is that just as the validity of a contract does not turn solely on the terms agreed to, but also on the fact of agreement, so justice consists in more than the proper distribution of rights and assets. Principles of justice should also be public, each of us affirming them in light of the fact that others affirm them too….Equally important is the *publicity* of its defining principles – that our reason for accepting them turns on others having reason to accept them too.' (Larmore 2002, p. 370, italics original).

[9] For the claim that consensual nudging respects personal autonomy, see Wilkinson 2013, p. 353.

[10] This invites us to think more closely about the relationship between choice architecture and public reason. Such reflection is sadly outside of the scope of this paper; however two points are worth mentioning here. First, the relationship between choice architecture, publicity and public reason may not be as strong as implied (as our publicity condition could take a non-Rawlsian form). Second, Rawlsians may object to choice architecture as a policy method that fails to treat citizens as free and equal (Rawls 2001, pp. 18-24; Rawls 2005, pp. 29-35).

of policy-makers must be open and scrutable. This transparency ensures that an individual is aware of the nudges that they encounter. As Bovens suggests, such transparency can take one of two forms: *type* and *token* interference transparency (2009, p. 216; see also Grüne-Yanoff 2012, p. 638; Blumenthal-Barby 2013, p. 191). Type interference concerns the form a policy might take. Token interference concerns which choice scenarios have been interfered with. These can be separated: I might know that a particular type of nudging is employed by an institution (type) but not know when and where I encounter it (token), or I might know that a particular choice situation is designed with a specific end in mind (token) but be unaware of the full range of nudges employed to achieve it (type).

Transparency compliments the second characteristic: to maximise avoidability, an intervention must be transparent in both ways. As with avoidability, it might be argued that transparency is in tension with effectiveness. By its very nature, the transparency condition reduces the opacity of permissible polices. If opacity is required for the success of certain nudges, then a tension may exist. Thus, a commitment to conditions of publicity and transparency is also a commitment to the idea that not all nudges need to be hidden to be effective.

Interpreted in this manner, I suggest that the normative guidance on the permissibility of interference contained within the most popular expression of choice architecture can rebut criticisms of heteronomy. So long as a nudge: a) is in line with a competent individual's authentically preferred ends, b) is easily avoidable, c) meets some form of publicity condition, and d) meets conditions of transparency, then that policy does not pose a threat to personal autonomy.

This is a high bar to clear, but I believe that the following types of policies meet these standards and thus offer choice architects an effective response to our concern:

i) Personalisable Default Rules – central to choice architecture, the idea that there is no neutral choice (*sans* any influence) is reflected in policies that

determine what should occur if an individual does nothing. Often relying on inertia, if such rules could be neither controlled nor avoided by the subject they would pose a substantial threat to her autonomy. However, if the default rules are personalisable (so that they can be shaped by the subject), and contain opt-outs (so that they can be avoided by the subject), then they can be designed to help the subject protect herself from errant preferences that she believes distract her from her pursuit of the good (Sunstein 2014b, p. 99).

ii) Choice Prompts – sometimes known as active choosing, choice prompts are a type of default rule that prompt an individual to choose (Sunstein 2014b, p. 95). So long as an individual is not cognitively overwhelmed by such prompts (Dworkin 1988, pp. 78-81) these policies can increase the number of opportunities for autonomous choice and may improve the subject's competency at such decisions.

iii) Framed Information Provision – the selective disclosure of information to the subject can be designed to improve her decision-making according to her own subjective standards (Sunstein 2014b, pp. 139-140). This may include providing specific information about particular means toward the subject's chosen end. The provision of information reduces epistemic costs for action and respects autonomy by engaging (rather than bypassing) the subject's will.[11]

4. Extending the Argument

In the previous section I argued that guidelines concerning respect for autonomy can be applied to choice architecture to validate some nudges as non-heteronomous. To make this claim, I characterised autonomy as a form of authentic self-rule. As this characterisation of autonomy is contestable, my argument will be strengthened if it could be shown as plausible over contrasting

[11] In contrast, framing effects intended to subvert the subject's will impose an epistemic cost on their behaviour. For more on the implications of this for consent, see Hanna 2011.

conceptions of autonomy. In this section, I seek to illustrate that my argument is sound according to accounts of autonomy that both accept and deny authenticity a central role. Further, I will extend my claim to argue that choice architecture is not only compatible with personal autonomy, it may also promote it.[12]

Those who accept authenticity as a core value of autonomy (as I have) tend to characterise autonomy as a process of motivational reflection to locate the authentic self, and then the privileging of authentically motivated actions over others. A seminal example of this approach is Frankfurt's (1971) account, which suggests that autonomous behaviour is best characterised by the agent's role in determining the relationship between their various volitions (i.e. effective desires that successfully motivate an agent to act). In its simplest form, his account relies on a hierarchy of desires (i.e. first-order, second-order and so on) which captures our capacity for motivational reflection and the sense of agential control that this capacity grants us. Frankfurt suggests that autonomous actions are those that follow from second-order volitions; desires about desires that successfully motivate us to act. Accordingly, autonomous behaviour consists in an agent successfully acting from a desire that he or she wants to want to act from. Frankfurt has since developed this account, and alternatives to his approach differ in the motivators they employ and the relationship they favour.[13] However each variant reflects the central thought of this approach: autonomy consists of a form of motivational self-reflection and control over the relationship between motivating phenomena.

From this perspective, choice architecture can respect autonomy because it need not undermine the agent's ultimate control over her motivations for action. Choice architects may structure a choice situation so that some reasons for action are easier to comply with (through personalised default rules) or more prominent (through prompting and framing). Each type of policy may make it

[12] For the stronger claim that nudging may be required to respect autonomous choice, see Sunstein 2014a.

13 See also Watson 1975; Young 1980; Dworkin 1988; Ekstrom 1993; Frankfurt 1998; Cuypers 2000; Bratman 2003.

easier for the subject to act in line with her hierarchy of volitions. Accordingly, each example of choice architecture can be validated as non-heteronomous according to this conception. Further, I suggest that we are also entitled to the stronger claim that choice architecture can promote autonomy in this form because nudges can overcome causes of unreflective motivation. For example, choice architects may provide relevant information, prompt an individual to reflect upon that information and design rules that, in turn, help her to choose according to her own conception of the good in situations where she would have previously done otherwise. In these instances choice architecture increases instances of authentic behaviour, thus promoting personal autonomy.

So on one popular account of personal autonomy, choice architecture can be argued to both respect and promote autonomy. Can the same be said for a conception of autonomy that eschews authenticity? In stark contrast to Frankfurt's progenitive account stand *relational* accounts of personal autonomy. These often emphasise the social and relational aspects of personal autonomy (e.g. Freidman 2003, pp. 15-19) as more important than internal reflection and motivational control (Oshana 2005; 2007; Garnett 2013; 2014).[14] Proponents argue either that personal autonomy is an inherently social capacity or that it is an individualistic capacity that requires a number of social conditions to be satisfied (Freidman 2003, p. 96). Either way the external conditions of autonomy dominate, especially our standing toward fellow agents.

Can nudging respect autonomy in this form? I believe that it can. The examples of choice architecture I have suggested do not require domineering relationships between policy-maker and subject, nor do they force the subject into a form of subjugation. Though nudging (like all policy-making) requires a power asymmetry, not all power asymmetries are dominating. Thus not all policies that rely on these asymmetries display inherently objectionable forms of domination. Specifically, the opt-out condition prevents this from occurring. So long as the policies can be avoided, the subject's opportunity to shape their life as they wish

[14] See also Meyers 1989; Mackensie & Stoljar 2000; Oshana 2006.

(authentically or otherwise) is not diminished. Further, the provision of information can help break down social stigmas, reduce distrust caused by misinformation, and undermine harmful stereotypes by increasing awareness of the similarity of other's circumstances. Such policies would do much to improve people's social conditions. Nudges may even reduce the scale and number of dominating relationships by broadening information networks and increasing instances of meaningful choice, thus promoting autonomy in this form as well.

5. The Epistemic Objection

So far I have argued that a subset of nudges can be validated as non-heteronomous and for this reason can respect the autonomous choice of agents. In this section I consider whether this claim is too simplistic. From the outset, my argument has relied on setting aside the question of motive to focus on assessing choice architecture as a means. I have suggested that i) personalisable default rules, ii) choice prompts, and iii) framed information provision, can each respect personal autonomy. A critic may agree with my criteria but object that the sorts of interventions that I propose fail to respect choice because they remain instances of a specific form of paternalism hitherto unconsidered – *epistemic* paternalism. Thus, even though the narrow range of nudges I have discussed are not strictly heteronomous in my use of the term, they still disrespect autonomy because they treat the individual as incapable of pursuing their own good. Specifically, they fail to respect the subject as a competent chooser by denying them the chance to make mistakes and be held responsible for the consequences. Reducing the risk of mistakes may appear prudent but could nonetheless infantilise individuals (e.g. Bovens 2009, p. 215).

The charge of epistemic paternalism is a pressing one for choice architects. Our tendency for imperfect reasoning is matched by our overconfidence in our ability to make correct judgements. Choice architects are alive to this fact and seek to respond to it (e.g. Rachlinski 2003; Glaeser 2006; Blumenthal 2013). Policies designed to protect us from our cognitive failings (particularly in judgements of risk) have been criticised as yet another form of paternalism. Epistemic

paternalism constrains an individual's methods of inquiry to improve her epistemic standing or facilitate her pursuit of veritistic ends (Goldman 1991, p. 118). Common examples of such interferences are mandatory standards for good conduct in scientific and legal reasoning. The filtering of information (e.g. through principles such as anonymity) is intended to focus our faculties to increase the likelihood of making correct judgements. As a recent advocate puts it: 'Epistemically paternalistic interventions are not designed to tell people what to believe, but how to come to believe things.' (Ahlstrom-Vij 2013, p. 95).

Because choice architecture seeks to harness our biases and improve our decisions, critics might suggest that it exhibits epistemically paternalistic tendencies, and that these tendencies fail to respect the choices of autonomous agents. This is particularly true for some of the policies I have championed as non-heteronomous, such as framing. The non-deceptive provision of information is a benchmark for respectful interaction because it does not attempt to undermine the voluntary and responsible conduct of the authentic will. Such interventions provide reasons to persuade an individual to act rather than coerce or manipulate them toward the same end; they intervene without interfering. Choice architects often frame information, and while the framing of information need not be deceptive, it is selective and intended to lead the subject toward particular outcomes. As such, it may constitute an epistemic threat to the autonomy of the subject.

Epistemic paternalism constrains our options concerning information collection. This relationship is important both for those who recognise personal autonomy as a reflective capacity and for those who recognise it as a particular standing toward others. It directly threatens what we might call epistemic independence or epistemic self-reliance – our ability to pursue knowledge and seek truth in our own way. For the epistemic objection to hold, choice architecture must undermine epistemic independence and epistemic independence must be a necessary condition of personal autonomy. If this is the case, then choice architecture necessarily undermines the epistemic conditions of personal autonomy.

To respond to this worry I will determine whether my examples of non-heteronomous nudges exert epistemic pressure and whether this pressure contravenes respect for personal autonomy. I suggest that although the answer to the former question is often yes, the answer to the crucial second question is no.

Let us consider the descriptive point first. Why might we think that the instances of choice architecture that I have characterised as non-heteronomous exhibit epistemically paternalistic characteristics? Let us take my examples in turn:

i) Personalised Default Rules – these policies need not apply to our collection and processing of information. But when they do, a rule will be epistemically paternalistic if an individual fails to personalise it, requiring the policy-maker to guide their decisions. The method of guidance will then determine the extent of the epistemic threat.

ii) Choice Prompts – unguided prompts may actually increase the risk of mistaken decisions (making them counter-productive as a form of epistemic paternalism). Again, the extent to which prompts are epistemically paternalistic will be determined by whether and how the subject's choices are guided by the policy-maker.

iii) Framed Information Provision – such policies could be cases of epistemic paternalism because the information is selected and presented by the choice architect in a manner intended to guide the subject's decision toward a particular goal in an unreflective manner. This may effectively distract the subject's attention away from certain pieces of information and toward others, thus constraining the subject's enquiry.

A version of each policy that I offered as non-heteronomous in §3 appears to threaten epistemic paternalism. If epistemic paternalism is objectionable in

terms of respect for personal autonomy, then the options for the choice architect seeking to avoid wronging the subject are drastically reduced. Determining whether this is truly the case requires us to consider the level of constraint involved and the importance of epistemic independence to personal autonomy.

Concerning constraints, let us again consider each policy in turn. As noted, a default rule will be epistemically paternalistic to the extent that the subject fails to personalise it. Personalisation is important to ensuring the first characteristic of good nudging. As such, a non-personalised default rule may be a bad nudge regardless. But crucially, the second characteristic of a good default rule, its opt-out, will prevent the rule from completely constraining (and thus undermining) the subject's epistemic independence. If the subject wants to proceed against a rule designed to reduce the risk of falsehood they will ultimately be able to do so. The rule can neither significantly change the costs of enquiry nor prevent the subject from opting-out. This ensures that even poorly designed default rules might not pose an objectionable epistemic threat. Well-designed default rules certainly should not. The same can be said for instances of guided or framed choice. These processes do not wholly inhibit an individual's ability to pursue the truth. Rather they draw her attention away from certain pieces of information to emphasise others. So long as those other pieces of information are available (possibly as a part of the choice architect's commitment to publicity) then the policies are validated. Making the subject work a little harder to access the information is unlikely to be a significant constraint; deceiving the subject by withholding that information is. As such, the epistemic problem closely mirrors that of heteronomy (Zagzebski 2012, p. 24). The response is similar – while intellectual dependence on others is compatible with self-rule, forced dependence isn't. Because nudges allow the subject scope for personalisation and the potential to opt-out, they cannot force the subject into epistemic dependence. Thus, they do not conflict with either of the conceptions of autonomy from §4.

I have argued that the suggested policies need not pose an overbearing threat to epistemic independence. Even if they (or other nudges) do, how important is

epistemic independence? It has been suggested that it is not a necessary condition of personal autonomy (Zagzebski 2012, pp. 24-26 & 250; Ahlstrom-Vij 2013, pp. 65-90). While I generally accept this point, the issue is complicated by the fact that this relationship depends on the content of the individual's goals. To some, epistemic independence will be an instrumental good. This is because the authentic pursuit of your conception of the good does not generally require you to independently learn everything you need for this pursuit. Rather, it requires you to learn various truths relevant to this pursuit, and you might better learn these truths though epistemically paternalistic acts.

However, some individuals will value epistemic independence in itself and orientate their conception of the good toward it (e.g. the scientific enquirer). This increases the likelihood of a conflict between paternalism and autonomy. But even in these cases some interference will be acceptable. Placing great intrinsic value in epistemic independence threatens to commit oneself the burden of learning even basic truths for oneself. This cannot dominate an individual's conception of the good as it conflicts with basic temporal constraints and more fundamental components of her conception that she would struggle to ignore (such as health, enjoyment, and personal relationships). Further, the relationship doesn't always hold. Epistemic independence sometimes competes with autonomy because it may lead you to fail to learn the necessary truths, leaving you ignorant of your options (Fricker 2006; Ahlstrom-Vij 2013, pp. 92-108).

Therefore, epistemic independence is unlikely to be of great importance to the many of us who are willing to be guided by or rely on others in our pursuit of our conception of the good. To those of whom it is of great importance, it cannot be of overbearing importance. Thus, epistemic dependence does not always contravene personal autonomy and, even when it appears to, some forms of dependence must be acceptable. Therefore, epistemically paternalistic interferences need not contravene personal autonomy, and accordingly the epistemic objection fails. Because of this, objections based on the supposed infantilisation of the subject must be based on some other ground.

6. Conclusion

Having refuted both objections, I conclude that choice architecture can respect the autonomous will and authentic choice of the subject. In defending this claim, I hope to have defused a potential objection on epistemic grounds, provided a set of necessary and sufficient conditions for choice architects to avoid the charge of heteronomy, and a compelling reason why policy-makers should avoid such policies in the first place.

Word Count: 5,814.

Acknowledgements

I would like to thank the editors and anonymous reviewers for their helpful comments on an earlier draft of this article.

Bibliography

Ahlstrom-Vij, K. (2013). *Epistemic paternalism: a defence*. London: Palgrave Macmillan.

Blumenthal, J.A. (2013). A psychological defence of paternalism. In *Paternalism: theory and practice*, ed. C. Coons and M. Weber. Cambridge: Cambridge University Press.

Blumenthal-Barby, J.S. (2013). Choice architecture: a mechanism for improving decisions while preserving liberty? In *Paternalism: theory and practice*, ed. C. Coons and M. Weber. Cambridge: Cambridge University Press.

Bovens, L. (2009). The ethics of nudge. In *Modelling preference change: perspectives from economics, psychology and philosophy*, ed. T. Grüne-Yanoff and S.O. Hansson. Heidelberg: Springer.

Bratman, M. (2003). Autonomy and hierarchy. *Social Philosophy and Policy* 20 (2): 156-176.

Conly, S. (2013). *Against autonomy: justifying coercive paternalism*. Cambridge: Cambridge University Press.

Cuypers, S.E. (2000). Autonomy beyond voluntarism: in defence of hierarchy. *Canadian Journal of Philosophy* 30 (2): 225-256.

Dillon, R.S. (1995). *Dignity, character and self-respect: essays on self-respect*. London: Routledge.

Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge: Cambridge University Press.

Ekstrom, L.W. (1993). A coherence theory of autonomy. *Philosophy and Phenomenological Research* 53 (3): 599-616.

Ekstrom, L.W. (2005a). Alienation, autonomy, and the self. *Midwest Studies in Philosophy* 39 (1): 45-67.

Ekstrom, L.W. (2005b). Autonomy and personal integration. In *Personal autonomy*, ed. J.S. Taylor. Cambridge: Cambridge University Press.

Ekstrom, L.W. (2010). Ambivalence and authentic agency. *Ratio* 43 (4): 374-392.

Feinberg, J. (1989). Autonomy. In *The inner citadel: essays on individual autonomy*, ed. J. Christman. Oxford: Oxford University Press.

Frankfurt, H.G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy* 68 (1): 5-20.

Frankfurt, H.G. (1982). The importance of what we care about. *Synthese* 53 (2): 257-272.

Frankfurt, H.G. (1998). On the necessity of ideals. In *Necessity, volition and love.* Cambridge: Cambridge University Press.

Freidman, M. (2003). *Autonomy, gender, politics*. Oxford: Oxford University Press.

Fricker, E. (2006). Testimony and epistemic autonomy. In *The epistemology of testimony*, ed. J. Lackey and E. Sosa. Oxford: Oxford University Press.

Garnett, M. (2013). Taking the self out of self-rule. *Ethical Theory and Moral Practice* 16 (1): 21-33.

Garnett, M. (2014). The autonomous life: a pure social view. *Australasian Journal of Philosophy* 92 (1): 143-158.

Glaeser, E.L. (2006). Paternalism and psychology. *The University of Chicago Law Review* 73 (1): 133-156.

Goldman, A.I. (1991). Epistemic paternalism: communication control in law and society. *Journal of Philosophy* 88 (3): 113-131.

Goodwin, T. (2012). Why we should reject 'nudge'. *Politics* 32 (2): 85-92.

Grüne-Yanoff, T. (2012). Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare* 38 (4): 635-645.

Hanna, J. (2011). Consent and the problem of framing effects. *Ethical Theory and Moral Practice* 14 (5): 517-531.

Hausman, D.M. and B. Welch. (2010). To nudge or not to nudge. *Journal of Political Philosophy* 18 (1): 123-136.

Hill Jr., T.E. (2000). *Respect, pluralism, and justice: Kantian perspectives*. Oxford: Clarendon Press.

Jolls, C. and C.R. Sunstein. (2006). Debiasing through law. *The Journal of Legal Studies* 35 (1): 199-242.

Kant, I. (2012). *Groundwork of the metaphysics of morals*. Cambridge: Cambridge University Press.

Kerstein, S.J. (2013). *How to treat persons*. Oxford: Oxford University Press.

Larmore, C. (2002). Public reason. In *The Cambridge companion to Rawls*, ed. S. Freeman. Cambridge: Cambridge University Press.

Mackensie, C. and N. Stoljar. (2000). *Relational autonomy: feminist perspectives on autonomy, agency, and the social self*. Oxford: Oxford University Press.

Meyers, D.T. (1989). *Self, society and personal choice*. New York: Columbia University Press.

Mills, C. (2013a). Why nudges matter: a response to Goodwin. *Politics* 33 (1): 28-36.

Mills, C. (2013b) The problem of paternal motives. *Utilitas* 25 (4): 446-462.

Oshana, M. (2005). Autonomy and self-identity. In *Autonomy and the challenges to liberalism*, ed. J. Christman and J. Anderson. Cambridge: Cambridge University Press.

Oshana, M. (2006). *Personal autonomy in society*. Hampshire: Ashgate Publishing.

Oshana, M. (2007). Autonomy and the question of authenticity. *Social Theory and Practice* 33 (3): 411-429.

Pi, D., F. Parisi, B. Luppi. (2014). Biasing, debiasing, and the law. In *The Oxford handbook of behavioral economics and the law*, ed. E. Zamir and D. Teichma. Oxford: Oxford University Press.

Rachlinski, J.J. (2003). The uncertain psychological case for paternalism. *Northwestern University Law Review* 97 (3): 1165-1226.

Rawls, J. (1971). *A theory of justice*. Harvard: Harvard University Press.

Rawls, J. (2001). *Justice as fairness: a restatement*. Harvard: Harvard University Press.

Rawls, J. (2005). *Political liberalism (revised edition)*. New York: Columbia University Press.

Rebonato, R. (2012). *Taking liberties: a critical examination of libertarian paternalism*. London: Palgrave Macmillan.

Sunstein, C.R. (2014a). Choosing not to choose. *Duke Law Journal* 64 (1): 1-52.

Sunstein, C.R. (2014b). *Why nudge? The politics of libertarian paternalism*. Yale: Yale University Press.

Sunstein, C.R. (Forthcoming). The ethics of nudging. Working paper available at SSRN.

Sunstein, C.R. and R.H. Thaler. (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review* 70 (4): 1159-1202.

Thaler, R.H. and C.R. Sunstein. (2008). *Nudge: improving decisions about health, wealth and happiness*. London: Penguin.

Trout, J.D. (2005). Paternalism and cognitive bias. *Law and Philosophy* 24 (4): 393-434.

Watson, G. (1975). Free agency. *The Journal of Philosophy* 72 (8): 205-220.

Watson, G. (2004) Volitional necessities. In *Agency and answerability: selected essays*. Oxford: Oxford University Press.

White, M.D. (2013). *The manipulation of choice: ethics and libertarian paternalism*. London: Palgrave Macmillan.

Wilkinson, T.M. (2013). Nudging and manipulation. *Political Studies* 61 (2): 341-355.

Young, R. (1980). Autonomy and the 'inner self'. *American Philosophical Quarterly* 17 (1): 35-43.

Zagzebski, L.T. (2012). *Epistemic authority: a theory of trust, authority, and autonomy in belief*. Oxford: Oxford University Press.