Cambridge Journal of Economics 2020, **44,** 33–54
doi:10.1093/cje/bez012
Advance Access publication 29 April 2019

'Better off as judged by themselves': a critical analysis of the conceptual foundations of nudging

Alexander C. Cartwright* and Marc A. Hight

Libertarian paternalism claims to differ from traditional paternalism by making people better off, 'as judged by themselves'. We argue that choice architects use 'better off, as judged by themselves' in a way that is systematically unclear and misleading. This unclarity, furthermore, makes recent debates about the efficacy and morality of employing nudges as public policy instruments in some cases are difficult, if not meaningless. Ultimately, the matter simply resolves into intuition pulling about values, making libertarian paternalism effectively equivalent to traditional paternalism.

Key words: Behavioural economics, Behavioural law and economics, Paternalism, Default rules, Nudge
JEL classifications: B49, D61, D69, D03

1. Introduction

In their influential work *Nudge*, Richard Thaler and Cass Sunstein argue for a form of what they call *libertarian paternalism* (Thaler and Sunstein, 2009). The idea is straightforward: since people are less than perfectly rational, they are prone to make decisions that adversely affect them. Because these deviations from perfect rationality are predictable, smart policy makers can craft policies to alleviate these adverse effects by (hopefully) inducing individuals to make better choices. Furthermore, *libertarian paternalism* claims to be able to produce beneficial outcomes without actually coercing anyone. Freedom of choice is preserved, hence no one is coerced, and many benefit.

Libertarian paternalism has sparked a large debate and received a variety of criticism. Hausman and Welch (2010) argue that many of Thaler and Sunstein's proposals are more akin to traditional paternalism while others are not paternalistic at all. Others consider the political economy of libertarian paternalism: Rizzo and Whitman (2009A) argue that libertarian paternalists face a knowledge problem), and Rizzo and Whitman

Manuscript received 30 October 2017; final version received 22 December 2018.

Address for correspondence: Alexander Chase Cartwright, Department of Management, Ferris State University, 119 South Street BUS 337, Big Rapids, MI 49307, USA; email: alexcartwright@ferris.edu
*Ferris State University (ACC) and Hampden-Sydney College (MAH)

(2009B) along with Wright and Ginsburg (2012) argue public policies that 'nudge' could easily establish a slippery slope that erodes liberty. Others highlight threats to privacy and agency more generally (see Kapsner and Sandfuchs, 2015). Other authors have highlighted potential ethical issues with 'nudge'-based policies (see Selinger and Whyte, 2011).[1] Libertarian paternalism has also come under attack largely on moral grounds as a variety of critics allege that employing nudges as a tool of public policy is ripe for abuse.

Thaler and Sunstein have resisted these criticisms. In his memoir, Thaler writes: 'Readers who manage to reach the fifth page of Nudge find that we define our objective as trying to "influence choices in a way that will make choosers better off, as judged by themselves"' (Thaler, 2015, p. 325). In other words, slippery slopes, knowledge problems, and other concerns are not relevant to their program since nudges make people better off as judged by themselves. Sunstein has responded to the moral objections to nudging by noting first that nudges are ubiquitous; they even occur naturally or spontaneously and cannot be avoided in any event. Second, he argues that given typical, universal, relatively uncontroversial values (e.g., welfare, autonomy and dignity) we are nonetheless well advised to add nudging to our public policy arsenal (Sunstein, 2015A).

Whether or not nudges are a moral policy tool is an important issue, but here we argue that the debate about the morality of nudging is premature. There are conceptual problems with employing nudges that must first be resolved before one can meaningfully engage in a debate about the morality of using them, much less a debate about their efficacy. In short, we argue that Thaler and Sunstein lack a coherent conception of how to *evaluate* the effects of nudging from the perspective of the individuals affected. As they themselves take pains to emphasise, their libertarian paternalism rests on the critical claim that individuals who are 'nudged' will be better off 'as judged by themselves' (Thaler and Sunstein, 2009, p. 5; see also Thaler 2015, p. 325).

What exactly it *means* to say that individuals will be better off as judged by themselves is not actually spelled out or discussed by Thaler and Sunstein, and only recently has Sunstein sought to clarify the criterion (Sunstein, 2018). Aside from a short piece acknowledging that the criterion is somewhat ambiguous (Sunstein, 2018) and a vague reference to a large monograph by Van De Veer (who defends qualified, moderate forms of paternalism) there is no analysis (De Veer, 2016). What we find instead, however, is an approach that systematically engages examples in a fashion whereby the authors help themselves to intuitive claims. One can frequently read how people are 'open to a nudge' (Thaler and Sunstein, 2009, p. 109) or even that they 'need nudges for decisions that are difficult and rare' (Thaler and Sunstein, 2009, p. 74). As Sunstein and Thaler put it, 'Public-spirited choice architects – those who run the daily newspaper, for example – know that it's good to nudge people in directions that they might not have specifically chosen in advance' (Thaler and Sunstein, 2009, p. 99). Yet exactly *how* such public-spirited individuals *know* that it is good to nudge is decidedly unclear, even if it has an intuitive pull. In this paper, we are not interested in arguing against the theory or its morality *per se*; we seek only to provide critical clarification of the theory and note some serious difficulties with the theory that are not obvious at first blush. Our point is to highlight that, although the theory advocated in *Nudge*

[1] For an extensive review of this literature, see Rebonato (2012).

sounds good on the surface, it hides some dangerous pitfalls for analysts and policy makers that heretofore have not been sufficiently discussed.

We start by providing an overview of the theory of nudges in the context of it being a form of libertarian paternalism. From there we engage in an analysis of how to assess the impact of nudges by examining what it might mean to say that nudged individuals are 'better off, as judged by themselves'. As a result of that analysis, we conclude that the debates about the morality of nudges are often misguided and rely on intuition pulling and ethical assumptions rather than informative argument. In turn, this result helps clarify certain disputes and problems with assessing nudge policies.

2. Libertarian paternalism and nudge theory

Some preliminary work outlining the core concepts of this discussion must be done, detailing, first, the concept of paternalism and, second, the nature of libertarian paternalism in particular. One standard characterisation of paternalism comes from Dworkin, who defines paternalism as 'the interference of a state or an individual with another person, against their will, and defended or motivated by a claim that the person interfered with will be better off or protected from harm'.[2] There are broader conceptions. One might define a paternalistic action as one that imposes a cost on one agent but brings (greater) benefits to society or a larger social group.[3] For our purposes here, we employ Dworkin's more traditional (and narrower) conception. When a parent forces a child to do her homework against her will on the grounds that she will be better off for having done so, that is a paternalistic act. Similarly, when a government enacts a law (with punitive provisions for non-compliance) requiring drivers to wear a seatbelt, the policy is paternalistic. The justification for the law is that people will be better off if coerced into more regularly using their seatbelts.

One initial difficulty with even this seemingly straightforward definition is the notion of 'interference', especially when applied to nudges. Assuming reasonably that interventions are a species of interference, the issue matters. What constitutes an interference with another person? At least two components seem necessary to classify an act as one of the interferences in the present context. First, the act must be deliberate. I cannot *unintentionally* engage in a paternalistic act, even if the outcome is otherwise the same. Second, interference broadly includes any kind of action intended to alter the behaviour or preferences of the subject. If I do something to you that shapes how you form your preferences in the first place (which, in turn, then alters your subsequent behaviour), that action constitutes paternalistic interference even if there is no immediate change in your behaviour. My decision to only allow my child to listen to martial music as she matures (on the grounds that only such music is conducive to the formation of good character) constitutes paternalistic intervention even if right now there might be no behavioural impacts. This decision of mine probably will shape her preferences in the future. The phenomenon is most commonly observed when applied to religious and political beliefs, where children tend to closely mirror the beliefs of their parents.

[2] See Dworkin (2017). Rebonato quotes a previous version of this definition from the same article (p. 21). The differences are not substantive and not relevant to the present issue.

[3] See Rebonato (2012, p. 21). These are often called 'pro-social' interventions as opposed to 'pro-self'.

Enter libertarian paternalism. Economists have traditionally employed the neoclassical model that appeals to axioms of rationality (as originally proposed by John von Neumann and Oskar Morgenstern and later generalised by Friedman and Savage, 1948) in which agents have preferences that are stable, consistent and context independent. Work in behavioural economics, however, suggests that there are systematic errors that people regularly make because they are human. To err is human; the traditional axioms of rational choice (stable, consistent and context-independent preferences) cannot and do not try to account for the systematic errors of judgment people can be expected to make. Thus, Thaler calls for a more 'human' economics to substitute traditional axioms of rationality in neoclassical economics with a theory of rationality grounded in experimental psychology.[4]

While the call to combine rational choice theory with cognitive psychology has spawned a great debate among economists (see McQuillin and Sudgen, 2012 for an overview of the difficulties in reconciling normative and behavioural economics), the implications of a systematically flawed rational choice model extend far beyond requiring economists to update their understanding of human choice and the models that reflect such choices. If humans systematically violate rational choice, they can consistently make themselves worse off while often having little or no power to improve. If this is true, there are major public policy implications. Systematic mistakes constitute a 'behavioural market failure' that, like other market failures, government can step in to alleviate. This line of thought has spawned a literature of research broadly termed 'Behavioural Law and Economics' and, when applied to the political process, 'Behavioural Political Economy'.

Thaler and Sunstein (2009) argue that because human beings exhibit systematic and predictable errors in judgment and decision-making, public policy should help people make what Thaler and Sunstein believe are better choices in contexts where humans are prone to error. They argue that every choice occurs within a context—an arrangement of alternatives specified in particular ways that they call 'choice architecture'. All choices contain some type of 'architecture' as they refer to context, and they contend that it would be impossible for a choice environment to be completely neutral about the way that alternatives are presented and arranged. Just as a building cannot lack architecture, neither can choices. Thaler and Sunstein suggest that insofar as someone must design the architecture of the choices we make, one could increase the individuals' welfare by improving choice architecture: the way choices are constructed and presented.[5] Libertarian paternalism is a systematic policy that aims to exploit the choices of individuals by altering the choice architecture.

Libertarian paternalism is, strictly speaking, distinct from nudging (Gigerenzer, 2015). 'Nudges are interventions that steer people in particular directions but that also allow them to go their own way' (Sunstein, 2015B, p. 515). In short, a nudge is an intervention that exploits predictable behavioural tendencies in individuals. Thus, a nudge is a *tool* for influencing individuals (both their behaviour and their preferences) without

[4] See Kahnemann (2013) for an overview of many of those advancements in behavioural economics and psychology.

[5] Although not explicit in *Nudge*, Thaler and Sunstein later note that the choice environment need not be chosen by a single individual; the choice environment may be the product of no single person's designs or intentions. As Sunstein writes, "even spontaneous orders and invisible hands turn out to nudge, sometimes in extremely important ways" (2015, p. 416).

using economic incentives or coercion. Libertarian paternalism is a *program* that employs nudges as tools, with the explicit aim of 'overcoming the unavoidable cognitive biases and decisional inadequacies of an individual by exploiting them in such a way as to influence her decisions (in an easily reversible manner) towards choices that she herself would make if she had at her disposal unlimited time and information, and the analytic abilities of a rational decision-maker (more precisely, of Homo Economicus)' (Rebonato, 2012, p. 6). One might employ a nudge in a particular situation without thereby advocating a general public policy that uses them to achieve well-defined ends. A nudge also need not be paternalistic (although Thaler and Sunstein admit to being paternalists[6]): one might attempt to influence behaviour in ways that do not necessarily benefit the recipient of the nudge. We restrict our discussion here, however, to the kinds paternalistic nudges advocated by Thaler and Sunstein, which are interventions that intend to promote the well-being of the individual nudged.

In many circumstances, consumers face a default choice or a default rule. That is, they face a pre-determined choice that can be changed but only if the consumer takes the initiative to opt-out or request differently. Default rules help consumers economize on the choosing process; that is, exert less time and energy to calculate an optimal decision given the consequences of a choice (Sunstein, 2015A). Thaler and Sunstein champion changing default rules as a type of relatively non-intrusive, welfare-enhancing choice architecture.

Changing the default rule to increase a chooser's welfare is the quintessential 'nudge' that Thaler and Sunstein advocate. According to them, nudges are so non-intrusive that they rightfully can be thought of as examples of libertarian paternalism. The core tenets of their libertarian paternalism are twofold: first, libertarian paternalism does not restrict or eliminate choice. A change in the default rule (default choice) effectively 'nudges' choosers without violating their autonomy. Second, and most important for this paper, the welfare analysis of libertarian paternalist policies is not derived or evaluated by the actor manipulating the choice environment, as it is with traditional paternalism. Thaler and Sunstein aggressively deny that they are traditional paternalists. They repeatedly remind us that they have no interest in telling others what to do (see Thaler, 2015, p. 325).

3. Better off as judged by themselves

To better understand the concerns about nudging, it is necessary to carefully unpack what the various claims mean. It is easy to overlook important details and employ unreflective concepts when dealing with complicated issues. And as Thaler and Sunstein complain about the use of abstraction when reasoning about nudges, we can use a concrete case to highlight the underlying problem.[7]

Consider a particular, concrete example of the paradigmatic scenario that Thaler and Sunstein use: the choice architecture concerning retirement investment (Thaler and Benartzi, 2004). Abe is a young man in his mid-20s, not long out of a Masters graduate program who has just landed a job at a company that provides a 401k retirement plan.

[6] 'Again: Thaler and I embrace paternalism, and so the AJBT [as judged by themselves] criterion is emphatically not designed to defeat a charge of paternalism' (Sunstein, 2018, p. 7).

[7] Sunstein even notes that 'Here as elsewhere, abstraction can be a trap'. (Sunstein 2015, p. 416).

Abe has to decide whether to save more money in a 401k now where the employer matches up to 5% of the contribution or keep the higher income that is immediately disposable. Now we want to know whether Abe should be nudged and in what manner. According to Thaler and Sunstein, we can easily craft choice architecture in a manner that benefits Abe 'as judged by himself'. In fact, we craft the company's retirement program to 'nudge' its employees in a direction that is good for them (as judged by themselves) by exploiting the well-known tendency to overvalue some present benefit that comes with a large, but future, cost. Such overvaluation is especially likely when the cost will be borne in the distant future (hyperbolic discounting), which could lead Abe to under-save relative to his lifecycle rate.[8] Such nudging is not coercive because it preserves choice and it makes him better off. Our choice architecture appears sound until one stops to ask a rather simple question: What does it *mean* to say that being nudged is in the best interests of persons *as judged by themselves*?

The claim that we are better off as judged by ourselves seeks to ground the discussion of welfare judgments in the subjective evaluations of the individuals involved. If one sets an independent, objective set of criteria, measuring welfare might still be difficult, but it is at least clear what success requires from a particular policy. When those judgments of welfare are grounded in subjective reports, there is a danger of confusing epistemological issues (how we can *know* the subjective states of individuals at any given moment) with ontological ones (*whose* judgments should we prioritise when we are discussing individuals who differ over time). Such a concern has led some scholars to argue that nudges might be best defended on political and *not* welfare grounds at all (Guala and Mittone, 2015). Sunstein, however, is quite clear that 'The lodestar is welfare, and under the appropriate conditions, people's own judgments are a good (if sometimes imperfect) way to test the question whether nudges are increasing their welfare' (2018, p. 2). We contend that conceptual muddiness in libertarian paternalism masks deeper problems when they are more carefully unpacked. What is required is a clear sense of what the standard 'as better off as judged by themselves' means given a clear set of welfare criteria and a method that enables us to reasonably make those judgments.

Sunstein has subsequently become aware that there is an issue here, but his early engagement of the problem is oriented only towards defusing 'ethical' (i.e., moral) objections. Sunstein writes

> Many nudges can be evaluated under the 'as judged by themselves' standard; consider a reminder, a warning, a default rule, or disclosure of relevant information. To see whether the standard is met, we would have to take each nudge on its own. But the standard will often provide sufficient guidance (2015A, p. 429).

We simply deny the assertion that the standard provides sufficient guidance, instead alleging that Sunstein is helping himself to an analysis that essentially begs the question. We need to know what the standard *means* before we can assess its appropriateness, and neither Thaler nor Sunstein provide that careful analysis, despite a recognition of the importance of the claim. Immediately after the above claim, Sunstein continues: 'If the choice architect is actually succeeding in making choosers better off by their own

[8] The rate at which one should save such that as income changes, one's consumption remains relatively unchanged over one's life.

lights, there would seem to be no objection from the standpoint of welfare' (Sunstein, 2015A, p. 429). Such a move misses the point. We are challenging Sunstein's ability to satisfy the antecedent of the conditional (viz. *if* the choice architect is succeeding…). One needs to know what success is and means before one may assert that choosers are 'better off by their own lights'.

When later pressed with concerns about identifying the individuals over time whose welfare we are supposed to be promoting, Sunstein seems to miss the force of the concern. Lecouteux argues that libertarian paternalism cannot coherently account for behaviours over time because its practitioners help themselves to the assumption that there are 'true preferences' of individuals over time (2015). He raises the reasonable possibility that it might not be irrational to discount future selves. What might matter most in welfare judgments is the ability to be the chooser of one's own preferences and identity. Sunstein responds by noting that there might be many ways to characterise what a person is (and is not) over time, concluding 'The question whether Oscar is the same person over time is not, in the end, a question of fact. It is an interpretative question…. The best way for Oscar, or for any human being, to make sense of his life is to understand himself as having a unitary identity' (Sunstein 2015A, p. 527). As the reply demonstrates, Sunstein seems to miss the point. Whether something is a good way for Oscar to 'make sense of his life' is irrelevant to the evaluation of the nudge. Sunstein argues that nudges *make people better off*—and nothing in his response indicates *how* he defends that claim.

Sudgen (2017) points out that Thaler and Sunstein apply the 'better off as judged by themselves' criterion inconsistently.

> One interpretation gives it a wide range of applicability, but drains it of its content as a repudiation of paternalism. The other interpretation gives more meaning to the claim that people want to make the choices they are being nudged towards, but applies to a much narrower range of cases than Thaler and Sunstein have in mind. The rhetorical success of *Nudge* depends heavily on this ambiguity (Sudgen 2017, pp. 115–16).

The first interpretation appeals to what Sugden calls an 'inner rational agent' model in which agents view the world through a shell or screen of cognitive biases. Thaler and Sunstein defend nudges as policy instruments on the basis of appeals to the *error-free* preferences of those being nudged. As Sugden argues, Thaler and Sunstein are essentially helping themselves to a model where the behaviours and preferences of individuals used to justify a nudge are simply assumed to be rational. Sugden's analysis concludes that nudging is traditionally paternalistic.

The second of Sugden's interpretations sees Thaler and Sunstein modelling agents as simply having akrasia (Sudgen, 2017, p. 119). Sudgen points out that simply because agents change their actions in light of a nudge is not prima facie evidence that they are better off. Individuals might well be worse off even when in certain cases they judge otherwise themselves. Consider the example of breast self-examinations (Gigerenzer, 2015, p. 377). Some choice architects wish to nudge women into performing monthly breast self-examinations even though there is little scientific evidence that so doing reduces cancer mortality. Indeed, the evidence suggests that programs designed to promote such monthly self-examinations leads to more biopsies, mammograms and other expensive procedures without any measurable benefit. Imagine a woman who, being nudged by an information campaign, changes her preferences as a result and now consistently performs these examinations each month. Sunstein himself notes that

nudging can often shape or 'construct' the preferences of individuals (2018, p. 7). If asked, the woman might well endorse the policy that nudged her, not having any sense of the underlying facts.

In a reply to Sudgen, Sunstein acknowledges the 'better off as judged by themselves' criterion is ambiguous in some cases, but he conflates the willingness to accept nudges with impacts on welfare—missing Sudgen's point. Sunstein advances a number of cases to show that, given people's antecedent preferences, nudging will make them better off as judged by themselves. A reminder sent to a forgetful patient to take his medication is a nudge that makes him better off as judged by himself. The example seems clear. But as the example of the breast self-examinations shows, we need a *mechanism* or standard by which we can make reasonable judgments about when and how to nudge. Maybe the patient is better off. Maybe the patient is not. It depends on the medication, his situation, the value being pursued and the nudge being employed. Listing examples is all well and good, but without a lucid, consistent and accessible set of guidelines to evaluate nudges, policymakers are left to rely on their own intuitions and values when crafting nudges. Our argument deepens this debate and seeks to add conceptual clarity: regardless as to how Sunstein and Thaler conceptualise agents or how acceptable citizens believe nudges are, Sunstein and Thaler's welfare standard is not conceptually clear.

Although we are not given any analysis of how to read the critical phrase, Sunstein does provide some examples that are meant to be intuitive guides. For instance, he writes, 'If a GPS steers people toward a destination that is not their own, it is not working well' (Sunstein, 2015A, p. 429; 2018, p. 3). We agree, but the example is not illuminating. What would count as *success* in the case of using a GPS? Since the goal is being made *better off*, we first need some standard to which one might compare one's welfare. The driving advice offered by a GPS is certainly a nudge (we may opt not to follow its advice, but it certainly impacts our behaviour). But note that even GPS devices often provide options for users. The route offered by the GPS might include fewer tolls, be more direct (i.e., less distance travelled), or be faster (by employing roads with higher speed limits). Those options suggest that users might employ different standards.[9] Furthermore, the person who employs the GPS might change later her mind, wishing she had taken a more scenic route, or had avoided high traffic areas. The ex post user's judgment might easily differ from that of the ex-ante user. Thus, the implication that a GPS is a case where users may *easily* judge whether they are better off by their own lights is simply misleading. It is misleading because intuitive examples cannot replace analysis. If we hope to avoid an endless series of back-and-forth exchanges milking intuitions from disparate examples, we need to explore the conceptual territory more carefully.

We suggest that Thaler and Sunstein need a standard (whatever that might be) that is lucid, consistent and accessible. The basic idea is that *any* appropriate policy needs to have a clear set of standards (lucidity), workings that are not self-defeating in virtue of producing contradictions (consistency), and have a clear mechanism for application (accessibility). This bar is meant to be minimal in the sense that *any* reasonable policy should be able to get over it. Nudges are policy tools. If the tool has no clear standard for determining success with respect to some end or value, then it is not possible to

[9] Additionally, a GPS is solving a fundamentally technical problem to which there is an objective 'optimum' solution; life choices are rarely as close-ended or as simple—they often involve several margins of trade-offs to consider.

effectively evaluate whether the use of nudges actually accomplishes its goals. If nudges permit contradictory outcomes, then since anything logically follows from a contradiction, there is no meaningful distinction among nudges. Knowing that nudges can be beneficial is idle if there is no way to know *which* nudges provide the desired outcomes. And there is an epistemic bar to get over as well: we must have some sense of how we know what the outcomes are, even when the standards are otherwise clear. The standard—lucid, consistent and accessible—is our attempt to define a basic threshold for evaluating whether a policy leaves one 'better off as judged by himself'. To the end of providing the needed conceptual analysis, we walk through three of the most obvious and reasonable possibilities of how to understand the critical phrase 'better off as judged by themselves'. We call them the *Simple Case*, the *Simple Future Self Case* and the *Objective Standard Case*. For the impatient, we argue that *none* of these cases can charitably be what Sunstein and Thaler mean. In other words, there is no coherent, consistent sense of the phrase being employed. None of the options can reasonably support lucid, consistent and accessible policy. In each case, the intuitive understandings attached to making intertemporal welfare judgments fall apart under analysis. The simple case does not produce an understanding of nudges that would require policy change in the first place. The simple future case, although it has initial intuitive appeal, fails to produce a theory that could support libertarian paternalist policies without betraying its own commitment to not giving preference to certain antecedent value claims. The final case collapses straightforwardly into traditional paternalism, undermining the libertarian goals of the project. As a result, we lack a critical conceptual component of the theory that in turn requires us to suspend judgment about whether employing a particular nudge is moral until more analysis is done.

Let's start with the simplest case and then engage the others in turn.
A) *Simple Case*: Person P at t_0 – the moment of decision – makes a judgment, and being nudged is, at that moment, judged by P to be in P's own best interests.

For Abe, the simple case states that, at the moment of decision, Abe judges that he should be nudged (presumably to participate in the 401k, but it does not technically matter in which choice direction he is being nudged so long as he judges that it is the nudging making him better off). The company might nudge Abe by making participation the default (and hence both easy and consistent with our human status quo bias). In the simple case, if Abe were to stop and reflect on the decision and the choice environment (regardless of whether he actually does so), he would judge according to his own values and preferences that being nudged (and making that decision) would be in his own best interest.

It is unlikely, however, that this is what Thaler and Sunstein have in mind, since by hypothesis Abe needs a nudge to alter his behaviour or encourage a particular choice. If at the time of decision (time t_0), he already judges that he should be saving and participate in the 401k program, then he does not *need* the nudge in the first place. If, alternatively, he judges that he should not save in order to keep the higher current income, then it is simply false that he judges himself to be better off by participating. Assuming he wishes to be better off, he is judging otherwise. The *point* of a nudge is to alter behaviour that, according to some standard, is deemed less than optimal. If that standard is the individual, then nudges simply do not make much sense. Thus, Thaler and Sunstein cannot reasonably be read as interpreting that people will be 'better off as judged by themselves' if we read that phrase as referring to the person at the moment

the decision is made. At best the simple case would be an instance of nudgers begging the question by assuming that people share their preferences at the time of decision. Attributing that reading to Thaler and Sunstein would be most uncharitable, hence we need a better understanding of the phrase.

One might object by arguing that the *Simple Case* reading is indeed non-sensical when it comes to nudges that involve changing the default rule, but the *Simple Case* might present an appropriate interpretation of the 'better of as judged by themselves' standard in other types of nudges. Thaler and Sunstein consider information disclosures a nudge. Examples might include a car's annual fuel cost appears on its window sticker, appliances that come stamped with annual electricity costs and restaurant menus that include calorie disclosures. The *Simple Case* might reasonably be applied to someone making a decision: one would be in a position to say that he would have made a different decision without the information disclosure. The immediate availability of the information might be considered a nudge.

We agree that the aforementioned cases involve influences on behaviour, but disagree that they are actually nudges in the relevant sense. There is little doubt that with more information consumers stand to make more informed decisions; this is true for both fully rational and less than rational agents. The need for more information is not grounded on any type of deficiency in one's rationality. Additional information might *influence* behaviour, but not all influences on behaviour are nudges. According to the theory, nudges are the result of crafting choice architecture in ways that exploit systematic tendencies to behave that are less than optimally rational. The simple presentation of information *by itself* does not meet that standard.

Consider an example that Thaler and Sunstein provide in their 'bonus chapter' listing additional nudges: the placing of carbon labels on products to reveal the carbon footprint of a variety of consumer goods (Thaler and Sunstein, 2009, p. 261). Perhaps it is independently laudable to make more information available to consumers, but the mere act of doing so is not actually a nudge *in the relevant sense*. What systematic tendency to err is being engaged in the example? Unless one believes that consuming products with larger carbon footprints is by itself a defect of rational behaviour, the answer is none. If individuals did not value the environment highly, or valued it less than other goods, then changes in behaviour as a result of the posting of the information might not be beneficial at all. The example *assumes* a value instead of leaving it up to the choosers as promised. There is no rational defect being nudged here. One might object by noting some consumers have free-floating intentions (not to harm the environment, for example) yet make consumption choices that are at some level incongruent with their stated values. Our response is twofold. First, this does not constitute a prima facie case of irrationality and, second, it still requires the choice architect to assume a value.

Perhaps, one might press the issue, contending that the defect lies in the inability to see the connection between one's own desires and a particular choice. In such a case, posting information about the carbon footprint might nudge the consumer by simply reminding her of that connection. Yet, there does not appear to be any principled difference between inattention, forgetfulness and such alleged defects (e.g., the inability to see connections). If the nudge theorist wants to claim such cases as example of nudges, then the scope of nudging has increased considerably. Governments would be justified in employing 'reminder nudges' all over the place in ways that seem uncharitable to the

position. Policies designed to alter behaviour based on the inattention to the state of one's shoelaces does not seem to be consonant with the spirit of Thaler and Sunstein's aims, even if one viewed such inattention as a cognitive defect.

As a result of the foregoing analysis, mere information disclosures, despite Thaler and Sunstein's affinity for them, often do not qualify as nudges by their own standards. Using the nudge welfare standard to evaluate a non-nudge would be misguided.

The most obvious and promising alternative to the *Simple Case*—and one suggested strongly by the prose used by Thaler and Sunstein—is that a future self would, reflecting back on the moment of decision, judge that being nudged was in their own best interest. This gives us a second possible reading.

B) *Simple Future Self Case*: Person P, at some future time t_n, judges at t_n that she is better off *now* (i.e. at time t_n) by having been nudged at t_0, regardless of P's beliefs at time t_0.

This particular case seems initially the obvious way to understand the claim, but it hides a great deal of complication. What could be more obvious than someone reflecting on a decision made in the past, and being glad (or dissatisfied) with that decision? Imagine that 25-year-old Abe chooses to participate in his company's 401k program. Thirty years later as Abe is planning in earnest for retirement, 55-year-old Abe is profoundly happy that his retirement nest egg is larger than it might otherwise have been. Since as a young man he might not have saved had he not been nudged in that direction, he judges that he is (now, aged 55) better off having been nudged.

The problem is that the simple future self case helps itself to a number of assumptions to which it is not entitled. The first assumption is that Abe aged 55 is unequivocally the numerically same person as 25-year-old Abe in all relevant respects and thus has the same set of values and preferences at 55 as he did at 25. In other words, even though Abe at 55 who was nudged and Abe at 55 who resisted that nudge both share an identity, they are in fact effectively different people: a possible self of Abe as his life might have been is not the same person as the Abe that was nudged.[10] Hence, a welfare comparison between the two of them is devoid of meaning. Although there are technical issues to be engaged here concerning the nature of judgments of diachronic numerical identity, for the purposes of this paper we will leave those concerns to the metaphysicians (Parfit, 1986; Velleman, 2008). We can easily reconstruct the problem in terms of the values and preferences of the individuals under consideration.

Imagine the following variant scenario. A young woman, let us call her Belle, is employed at the same time as Abe at the same company. At age 25, she is thus offered the chance to set aside money in a 401k where her employer will match the first 5% of her contribution. But Belle has just been scraping by through college and graduate school and has amassed considerable debt. In addition, she values material acquisition, especially since many of her friends have things and opportunities she does not. She thus decides at time t_0 (the moment of decision) not to divert a part of her salary into retirement savings, opting to alleviate her debt and have more discretionary income. Many years later, Belle's debts are all or mostly paid and she has acquired many things (cars, electronics, vacation experiences etc.). Now, at age 55, she examines her retirement

[10] It is worth emphasising here that the issue is not identifying persons over time, but comparing welfare between one person and another person who might have been.

portfolio and laments the decision not to participate. She wishes she had invested in her retirement when she started at age 25.

The problem is that Belle's values and preferences have substantively changed. It is inappropriate to claim that Belle judges that she herself (at time t_n, say at age 55) would have preferred to have been nudged at t_0, because the perspective from which the claim is being made already includes the advantages of spending the extra income. Belle is not making, and probably cannot make, an appropriate comparison, because she cannot reasonably know what her life would had been like had she not had the extra income, and hence not paid down her student debt and not enjoyed a few extra material comforts. In a sense, to ask otherwise is to ask her to compare herself, her values and her choices with another individual who, while intuitively similar in many respects, is not actually her.[11] Of course, Belle is going to report that she would have preferred the nudge when aged 55, but that is not the standard by which one can reasonably assess whether she would make the same decision. In effect, the simple future self case results in nudges that *beg the question* in favour of the nudge by giving preference to Belle's values at a time future to the moment of decision.

Now consider the example with a different decision. Imagine that Belle, aged 25, elects to save for retirement and take advantage of the 5% company match. Thirty years later, at time t_n we now ask whether Belle would judge then (at t_n) that she is better off. We face a similar problem. At t_n Belle now has the advantage of having saved more and has a larger retirement pool, certainly a good thing in many respects. That benefit or value is easily identifiable and present. But she has no reliable way to assess the opportunity cost of the decision (we here assume, reasonably we think, that *no* decisions are free of opportunity cost, so the details of the particular case are not vital to our point). What might she have done with the money had she saved instead of spent? It might have gone to material goods, or to an alternative investment that might have paid more handsome dividends (catering to two distinct values). Furthermore, this Belle probably still has debt, which likely lowers her net worth. In short, Belle has no way of making any reasonable judgment. Hence, in comparing a clear benefit with two costs that she cannot fathom let alone articulate, of course she will judge herself better off.

Oddly enough, the example highlights an ironic upshot. It might well be the case that 55-year-old Belle is making a predictably irrational judgment about the decision made when she was 25 years. She overvalues her present benefits and fails to appreciate the opportunity costs of that decision. It is empirically possible—both in terms of Belle at 25 years and Belle at 55 years—that given her values (respectively, at each age, even if they differ) her judgment that she is better off having saved is mistaken. The mistake might also be a feature of our flawed decision-making processes. Nudging does not guarantee that decision-making will improve because it does not provide a coherent mechanism for making the intertemporal comparisons.

To make matters worse, behavioural psychology has shown that there are many possible biases at play. Loss aversion (one implication of prospect theory) suggests that losses 'hurt' more than gains, such that people will favour certain gains over uncertain gains that are larger (Kahneman and Tversky, 1979). Part of the point of nudge theory is to manipulate the predictably irrational tendencies of human agents, so it

[11] Note that Belle's perspective is limited not only by metaphysical and epistemological concerns, but also by cognitive biases.

will not do to ignore them when they complicate the evaluation of the fruits of various nudge policies. Loss aversion is but one example of systematic bias that makes these judgments over time problematic. People favour judgments that make themselves look better, confirm their own views and preserve the status quo (to list a few). The upshot is clear: if we judge 'better off as judged by themselves' to mean simple future self case judgments, then it appears that Thaler and Sunstein are begging the question. They are helping themselves to a presentation of the cases that favours their own analysis by giving preference to the values they wish to highlight (or, to be more precise, giving preference to the individuals at times where they are more likely to have the values the nudger wishes to emphasise) and have the nudge policy assessed accordingly.

When evaluating the welfare impact of their 'Save more Tomorrow' retirement savings nudge, Thaler and Sunstein do recognise this complication. The pair responds that since few employees opted out of the savings plan after having been nudged into the plan, the plan obviously benefited them. They argue that revealed preference is (more) acceptable as a welfare criterion if people have a cooling-off period allowing their system II cognition (the more reflective deliberate side) to influence their opinion (Thaler and Sunstein, 2003, p. 1191).

This response, however, does not engage the conceptual problem explained above, nor does it absolve them from the charge of incoherence. Thaler and Sunstein previously argued that, "What people choose often depends on the starting point, and hence the starting point cannot be selected by asking what people choose" (Sunstein and Thaler, 2003, p. 1191). If the starting point determines what people choose, then the fact that people are afforded a cooling off period is a red herring. If starting points always influence revealed preference and someone must choose a starting point (as Sunstein and Thaler repeatedly argue), then revealed preference and welfare can never be known to align, even though they need them to align in this case. If Sunstein and Thaler are able to determine when revealed preference is an acceptable standard and when it is not, they should explain how that is done. Without such explanation, it appears as if revealed preference becomes an acceptable standard only when the preferences revealed align with those of Sunstein and Thaler—hardly an appropriate outcome.[12]

One challenge to our analysis here is that our criticism might prove to be too strong. One might wonder whether it is possible to do *any* intertemporal welfare comparisons if we are correct. That appears to be an absurd result, since clearly welfare policies are designed to do exactly that—make people demonstrably better off than they were previous to the implementation of the policy. Such an objection, however, misses the important detail of *where* the problem lies. Making intertemporal welfare judgments is not at all problematic when employing an external or objective standard. Evaluating a policy based on its ability to raise income or improve educational levels presents no special difficulties. Only when the standards being invoked rely essentially on subjective judgments do the problems arise. One can stipulate welfare criteria and make measurements in light of them when those standards are independent of the subjects, but this is not straightforwardly possible when the very nature of the standard being

[12] Intuitively, one would expect welfare gains from a change in the default rule to be highest when choosers are relatively homogenous; this is hardly the case when it comes to personal finances. Furthermore, Zywicki (2017) illustrates that the need for such a nudge is ungrounded as there is little empirical evidence that workers systematically under save for retirement.

employed is subjective and hence subject to variances for which one cannot account. Nothing in our analysis makes social science impossible; we are alleging that theories like revealed preference and standards like 'better off as judged by themselves' are significantly more susceptible to errors such as begging the question (with respect to the values and standards by which the theories are evaluated).

Though behavioural economics positions itself as opposed to traditional neo-classical economic analysis, another thoughtful retort to the challenges we raise to using the *Simple Future Self Case* of 'better off as judged by themselves' might involve using part of the neoclassical methodological toolkit: the validity of a theory's assumptions is unimportant so long as the theory has predictive power. In other words, behavioural economists might grant all of the objections above, yet argue that so long as people behave *as if* they are judging themselves to be better off as described in the *Simple Future Self Case*, then whether or not actors actually make coherent comparisons is irrelevant since the theory yields robust predictions. As an example, advocates of this response might describe a skilled truck driver. He is able to drive the truck backwards, around tight corners and through harsh driving conditions with relative ease. He has these skills even though he has no knowledge of Newtonian physics; yet we can, nonetheless, assume that he has this knowledge of physics and perfectly predict his behaviour (Friedman, 1966). One might say that the driver behaves 'as if' he knew the physics, even though he lacks the specialised knowledge. Thus, even though Belle cannot coherently make a welfare comparison at age 55 to judge whether she is better off for having participated in her 401k program, if she judges so anyway, then the policy is, nonetheless, vindicated and such nudges will be justified. She is behaving 'as if' she had the ability to make such intertemporal comparisons.

The 'as if' criterion, however, does not actually avoid the concerns associated with the *Simple Future Self Case*. In economics, the 'as if' defence of false assumptions is often applied to cases which, while they might never be true (i.e., there will never be perfect rationality just as there will never be a perfectly frictionless plane), are nonetheless logically consistent and falsifiable. Those requirements assume the stability of the assumptions. As we have already noted, we have no guarantee of the stability of the preferences and values of the individuals being nudged. Given the real possibility that preferences and values change, we are, in effect, privileging the supposed values of older persons who are nudged when younger. And since we cannot be certain what values will change or how, nudging turns into simply selecting values the choice architect favours to guide policy.

If we decide to select a particular value held by an individual at a particular time, we face additional concerns. Whether Belle judges at age 55 that she is better off is no longer strictly the primary concern. We are not trying to *predict* her behaviour; we are trying to implement policies that correct predictable irrationalities and in fact make her better off. Individuals might also judge that they are better off precisely because they are under the influence of some cognitive pattern of thinking that is sub-optimally rational. In other words, if people can systematically err about decision-making generally, they can systematically err about judgments concerning their own well-being. When it comes to nudges interpreted as simple future self cases, whether she thinks she is better off at some future date to the decision is not the point. Whether she is in fact better off is the critical issue. Hence, we have no principled way to ascertain whether the subjective judgment in one case—when compared against other possible

judgments—is in fact correct. Thaler and Sunstein would not, we assume, grant that those people should not be nudged merely because they *judge* (mistakenly) those decisions to have been right.

The libertarian paternalist might raise a new objection at this point, claiming that a reasonable case can be made for nudges in a society where there is broad agreement about the values and preferences being employed when implementing nudges. So long as (a) choice is preserved and (b) the values being invoked are widespread, it is reasonable to simply adopt those values and apply them to individuals, no matter the time reference. Such a move promises to avoid the concerns about subjective intertemporal welfare comparisons. This strategy gives rise to what we call the objective standard case.

C) *Objective Standard Case*: Person P, at some future time t_n, according to a value standard independent of P, would judge that she herself is better off having been nudged at t_0, regardless of P's beliefs at time t_0.

The basic idea is that regardless of what P happens to believe or value at t_0, there is some reason to think that P should evaluate the appropriateness of being nudged according to some objective standard (or at least a standard that is independent of P, e.g., some 'perfect' conception of the good or a societal standard). Even if P is not capable at t_0 of performing that analysis, if she were, she would agree to be nudged. The critical difference between this case of being nudged and outright paternalism is that P retains the option to choose otherwise, even though doing so might (likely *will*) require more effort, work or cost.

The trick here, of course, is to pick the 'right' independent standard or value. Yet this case essentially reduces the nudge architect to a straightforward traditional paternalist unless P freely chooses the values or preferences in question. In this case, the choice architect, a person in a position of power and authority, decides to choose a value to nudge one towards by exploiting his cognitive biases. If the 'right' values are not adopted, then it does not matter what the outcome is, since the case is one of what is 'best' for P independently of P's desires or values. If P happens to value whatever value has been 'objectively' selected, so much the better, but that does not change the basic fact of the case that P's desires are not really relevant. Sunstein explicitly denies that the 'better off as judged by themselves' standard appeals to such objective standards (referring to what he calls 'perfectionism'), and rightly so.

Initially Sunstein appears to see exactly where there is a deeper problem, noting that *when* we ask a person to make a judgment matters. 'Choosers' ex ante judgments might diverge from their ex post judgments' (Sunstein, 2015A, p. 430). Just so; this is precisely one of the concerns we emphasise above. Yet when confronted with the problem, Sunstein appears to think that the concern is a minor one, perhaps resolved on a case-by-case basis through empirical analysis. His response to the problem is illuminating.

But when ex post and ex ante judgments differ, the standard becomes more difficult to apply. [A] One option would be to use active choosing to see what people actually want. [B] Another would be to explore the number of opt-outs under different default rules. [C] A third would be to attempt a more direct inquiry into people's welfare under different forms of choice architecture, though admittedly any such inquiry raises challenges of its own (Sunstein, 2015A, p. 431).

We have inserted our own markers in square brackets to aid discussion. To see that Sunstein has fundamentally missed the primary problem, just consider [A] carefully.

Active choosing would not resolve the problem. The goal is to understand how to assess the impact of a particular nudge on an individual as judged by that individual. Appealing to active choosing does not provide any guidance at all. Because the judgments of a person can and do vary over time, there is no standard by which to make a consistent judgment. In fact, [B] and [C] betray the same kind of confusion about the problem. Exploring different (kinds of) opt-outs and different default rules, or employing direct inquiries does not resolve the question of *whose* welfare and preferences one is engaging.

Sunstein further seems to write as if the problem is simply converging the values, preferences and judgments of persons over time. If the ex-ante judgment were the same as the ex-post judgment, then perhaps all would be well. Yet that forgets the point of the nudge, which is to modify the behaviour of individuals *away* from the less-rational patterns normally found. If Belle at t_0 made the same judgments as at t_n and preferred to invest in her 401k, then the nudge would not have been required in the first place. But if her judgment at t_n matched a sub-optimally rational judgment at t_0 not to participate, then it is difficult to see how anyone could reasonably claim that Belle is better off by her own judgment. In either case, the analysis lapses into incoherence.

4. Welfare Comparisons Amongst the Irrational

Perhaps the solution for advocates of nudging lies in further behavioural economics. There is a new, but relatively extensive literature on behavioural welfare economics that attempts to tackle the problem of making welfare claims when consumer choices are inconsistent.[13] Bernheim provides an excellent summary of the issue at hand:

> According to one interpretation, standard normative analysis respects the decision maker's true objectives, which her choices reveal. Because the behaviors of interest by definition defy conventional rationalizations, that interpretation requires one to entertain unconventional rationalizations. But as a general matter one can offer many unconventional rationalizations for any particular behavioral pattern. Thus, knowledge of a choice correspondence may shed insufficient light on objectives, and hence on the mapping from the objects of choice to well-being. One can attempt to identify welfare either partially or completely by restricting the set of allowable unconventional rationalizations, but useful restrictions are difficult to justify. Those conceptual difficulties have led some to argue that economists should try to infer well-being from self-reported happiness and/or neurobiological activity. Unfortunately, it is every bit as problematic to identify useful information concerning internal well-being from such data as it is from choice. (Bernheim, 2009, p. 315)

Some authors have developed methods for approximating welfare under scenarios where choices are not consistent. The best-known and most notable approach is laid out in Bernheim and Rangel (2009). The authors suggest replacing the standard revealed preference relation with an unambiguous choice relation: 'roughly x is (strictly) unambiguously chosen over y iff y is never chosen when x is available' (53). When applied to choice scenarios, this criterion sets bounds on our welfare conclusions. Therefore, if an irrational chooser selects a set of intransitive choices, economics might be able to make some statement about his welfare.

[13] For an overview of the debate on how to assess welfare without revealed preference, see Bernheim (2016, 2009), Gul and Pesendofer (2007), McQuillin and Sudgen (2012), Reidl (2010), Atkinson (2011), Fluerbaey and Schokkaert (2013), and Rubinstien and Salant (2012).

An example here might help illustrate the point. Abe periodically receives salary increases. Every time he receives a salary increase, he has the option of spending his additional wealth on a Caribbean vacation, but he never does, instead choosing to save and spend his money on other things. Abe might not make perfectly rational or even consistent (transitive) decisions about how to spend his money, but the fact that he never spends his additional income on a Caribbean vacation does tell us that he values the Caribbean vacation less than his other options, even if those other options are inconsistently chosen.

An obvious problem here is that this approach does not help us draw welfare conclusions about inconsistent choices, and the model requires the choice architect to do a potentially impossible (and at least controversial) task: the model requires the choice architect to determine which elements in the chooser's environment are irrelevant to a consumer's choice (i.e., which elements are 'distortive'). In other words, the choice architect gets to determine which cognitive deficiency is impacting the chooser's decisions and based on that, determine which elements of the choice environment (i.e., options) are (allegedly) distorting his choice.

As has been discussed above in the analysis of the simple case, nudges would only be necessary where the chooser could not herself identify the distortive element. It is unclear how a choice architect would determine which elements are in fact distortive given the inherently complex and subjective nature of value and choice. For example, Abe's decision not to save for retirement could be motivated by a desire to pay down debt, purchase a new car, save for a new home or any of a host of other reasons. Without taking Abe's revealed preferences at face value, the choice architect must determine what Abe's 'true' desire actually is and thus must make a judgment about which element has distorted his actual choice from his 'true' desire. Furthermore, in the spirit of Alchian (1950), where Alchian points out that it is impossible to define profit-maximising behaviour ex-ante, it is arguably impossible to identify distortive mechanisms prior to making a choice and incurring the consequences.[14]

The fact that throwing revealed preference out the window requires the choice architect to determine which element is distortive is not only consequential because it is difficult to know a chooser's true intentions, but also because allowing the choice architect to determine which elements are distortive effectively allows the choice architect to determine the ultimate choice outcome.[15] For example, Abe receives a raise and decides not to increase the contribution to his underfunded 401(k) plan but instead decides to put the additional income towards a car payment on a new car. A choice architect might reasonably assume that Abe is falling victim to hyperbolic discounting by vastly overestimating the value of the new car while simultaneously underestimating the power that an extra few hundred dollars a month could do to his retirement portfolio. After all, the car will depreciate, require maintenance, increase his insurance payments and end up decreasing his disposable income by much more than the car payment itself, whereas a small increase in contribution to his 401(k), with compound interest, could increase his lifetime wealth by tens of thousands of dollars. However,

[14] Clippel (2014) presents another method for making welfare claims when choices are inconsistent, but again, the method does not eliminate all ambiguity and places a high epistemological burden on the choice architect (in this case requiring the choice architect to know choosers' "willpower index" pp. 22891). For an additional discussion, the application of Alchian (1950) to behavioural economics, see Manne and Zywicki (2014).

[15] For a formal model on this point and extended illustration, see Manzini *et al.* (2011).

Abe could be in dire need of a new car and the car he bought could have been a fantastic deal. Yet a nudge towards additional savings will result in additional savings, just as the aforementioned empirical studies on nudges for 401(k) accounts show.

The broader point here is that a single decision (e.g., to increase one's savings rate or not) is embedded into a complex and highly individualised opportunity set. The sheer complexity of the margins in which costs and benefits are traded off (even setting aside notions of subjective value) combined with the fact that different people will fall prey to different cognitive deficiencies, make it highly unlikely that a nudge on all choices will improve welfare. If there is no systematic way to measure how welfare changes via nudges, then nudging—to the extent it does change behaviour—appears to change behaviour primarily in ways consistent with how choice architects decide to nudge them.

Bernheim *et al.* (2015) illustrate how widely potential nudges can vary when choice architects have imperfect knowledge: in attempting to devise a welfare enhancing default rule for 401(k) contributions, the optimal savings rate ranged from the maximum amount an employer will match to zero. The appropriate default rule varies from one extreme to another because it depends on the particular behavioural bias impacting each individual and the scope of the choice domain deemed welfare relevant among other factors. The authors show how, depending on these factors, the default contribution rate could be the highest matched by the employer or zero. In other words, the optimal savings rate for Abe could be the maximum contribution or zero depending on the assumptions we make about (A) his cognitive deficiencies and (B) how many margins we include in his choice.

Under what circumstances would one reasonably conclude that a 0% savings rate is optimal for Abe? Examples are easy to generate. Perhaps, Abe has a significant amount of high interest rate debt, such that it would be best to pay down that debt before saving. He might have family obligations (such as an ill child or a family medical emergency) where the ready availability of funds is more important to Abe than long-term saving. Abe might have a partner whose employer's retirement contributions are more generous his own, thus making it reasonable to shift how they save while preserving enough income to pay the bills. Or perhaps Abe is independently wealthy from having inherited substantial wealth. Additional savings in that case might provide him near zero marginal utility. With greater disposable wealth, Abe might have access to investments that are more lucrative than his 401(k) can provide. In short, there are plenty of reasons why the optimal savings rate might be zero.

The optimal rate might also be impacted by how many margins we include. If Abe values mountain climbing and recognises that he needs to be young and healthy to indulge that activity, he might save less in order to climb more. Such a decision would be rational and optimising on those margins. It all depends on what Abe values and how he orders those values. All of these concerns are more worrisome when we remind ourselves that even choice architects have imperfect knowledge. The confidence that choice architects display in making arguments for policies seems to be more a reflection of their own values and biases than an accurate accounting of the values of those impacted by the policies.

Additionally, Rubinstein and Arad present experimental evidence showing how the nudge method can have negative welfare effects (2017). Specifically, participants prefer not to be nudged; the authors argue that the chooser's preference not to be nudged should be considered as part of the overall welfare impacts nudges have (p. 22).

Therefore, even if the objective standard case was the correct interpretation of the 'as judged by themselves' standard, and this standard was analytically tractable, Sunstein and Thaler would also need to consider welfare consequences cause from the nudge process itself.

There are two important upshots from this analysis: first, there is not a cohesive behavioural welfare economics, so even assuming that Sunstein and Thaler's 'better off as judged by themselves' standard should be interpreted as we have in the objective standard case, they do not have a basis for making conclusions about welfare. Second, attempts to circumvent our earlier challenge consistently reduce to traditional, but hidden, forms of paternalism. Trying to make intertemporal welfare comparisons coherent by smuggling in objective standards in the evaluation portion makes the view traditionally paternalistic after all. We suggest that Thaler and Sunstein are misguided in thinking that they are proposing a form of 'libertarian' paternalism, but not because the view is not libertarian or because it is ripe for abuse. Instead, we think that they are fundamentally unclear about how to evaluate the libertarian portion of their policy prescriptions and are thus *masking* the fact that they are smuggling values and preferences into the analysis. Our effort here, to be clear, makes no attacks on paternalism or its viability; we only want the theory and its consequences to be clear.

Consider the following example. In recent years, everyday purchases are more commonly made electronically; as debit card use has increased so have the number of checking accounts that accompany them. When a consumer makes a purchase for more than his available account balance, the bank may cover the charge but impose a fee for overdrawing the account; the fee is commonly referred to as an NSF (non-sufficient funds) fee. Low income and disorganised consumers might rack up hundreds of dollars annually in NSF fees in exchange for the convenience of not having to make financial plans. In 2009, the Federal Reserve and Consumer Financial Protection Bureau (CFPB) helped institute a new policy changing the default rule for new checking accounts. Post 1 July 2010, consumers have to opt-in to have overdraft protection, and by default checking accounts are opted out of overdraft protection but, therefore, those account holders will never have to pay an NSF fee.

This change in default rule represents a classic nudge like the ones that Sunstein and Thaler advocate, but the nudge also lends itself to being an example of making customers better off via an objective standard. Quite reasonably, consumers might want to minimise the fees they incur and this nudge could certainly do that.

The issue is that even if a consumer consistently and accurately states that he wishes to minimise the NSF fees that he pays, this does not reveal an unwillingness to pay. That is, we do not know at 'what cost' a fee might be worth incurring. Every time an NSF fee is charged, the consumer faces a trade-off, the specifics and nuances of which the libertarian paternalist does not know and cannot account for by changing the default rule. Even if we could elicit an objective value from a consumer, without fear of changes to their ex-ante and ex-post preference changes, the willingness to pay to realise that value will always be context specific. In other words, even if Abe always agrees with Thaler that Abe wants to minimise his expenses on NSF fees, he does not articulate a willingness to pay in order to minimise his NSF expenses; Abe does not articulate the circumstances under which he wishes to minimise his NSF expenses. In fact, it is doubtful that Abe even knows all such circumstances, let alone communicates

them to the libertarian paternalist who is even less able to code them into the perfect algorithm.

Furthermore, by not considering tradeoffs, nudges based on an objective value also favour a presentation of the cases that gives preference to the nudge and values that the nudgers wish to highlight: changing the default such that consumers must opt-in to NSF fees resulted in only 44% of frequent overdrafters choosing to opt-in. Hence, on net, the average amount of total fees levied was reduced, yet we have no tractable way of evaluating whether or not consumer welfare improved. To call a reduction in NSF fees imposed an improvement or success for consumers is to substitute the regulator's value for that of the consumers—and this is in effect an example of traditional paternalism.

5. Conclusions

Nothing we have said thus far indicates that we ought to seek or avoid policy prescriptions that nudge; we are simply noting that the standard invoked by Thaler and Sunstein themselves is inadequate. Whether a particular nudge is moral or appropriate depends on the standard one uses. That standard and how it works needs to be lucid, consistent and accessible.

We have intended to engage Sunstein and Thaler on their own grounds by evaluating libertarian paternalism based on whether it makes people 'better off as judged by themselves'. However, those who advocate improving the world via nudges rest their efforts on a welfare criterion that is far from fully specified. We have analysed the three most charitable interpretations of 'better off as judged by themselves' and conclude that all three are far less than lucid, consistent or accessible.

It is hard to imagine how Thaler and Sunstein could revise the 'better off as judged by themselves' welfare criterion in a way that both distinguishes their program from traditional paternalism while also eliminating the concerns we have raised. It seems that any attempt to do so would require an appeal to revealed preference. Yet if irrationality exists, then revealed preference is sometimes 'wrong'. If choice architecture really is omnipresent, then we cannot fathom how libertarian paternalists could determine or legitimise when it would be acceptable to rely on it.

Without clear welfare gains, libertarian paternalists must re-evaluate or re-specify their welfare criteria. If their program is to continue in its current form, Thaler and Sunstein should admit to being regular paternalists since this is what any workable interpretation of their welfare criterion reveals under scrutiny.

Bibliography

Alchian, A. A. 1950. Uncertainty, evolution, and economic theory, *Journal of Political Economy,* vol. 58, no. 211–221

Atkinson, A. B. 2011. The restoration of welfare economics, *American Economic Review,* vol. 101, no. 157–161

Bernheim, B. D. 2009. Behavioral welfare economics, *Journal of the European Economic, Association,* vol. 7, 267–319

Bernheim, B. D. 2016. The good, the bad, and the ugly: a unified approach to behavioral welfare economics, *Journal of Benefit-Cost Analysis,* vol. 7, 12–68

Bernheim, B. D., Fradkin, A. and Popov, I. 2015. The welfare economics of default options in 401(k) plans, *American Economic Review,* vol. 105, 2798–2837

Bernheim, B. D. and Rangel, A. 2009. Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics, *Quarterly Journal of Economics,* vol. 124, 51–104
Clippel, G. 2014. Behavioral implementation, *American Economic Review,* vol. 104, 2975–3002
De Veer, Van. 2016. *Paternalistic Intervention: The Moral Bounds on Benevolence, Princeton University Press*.
Dworkin, G. 2017. 'Paternalism' , in Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy*, Winter 2017 Edition, https://plato.stanford.edu/archives/win2017/entries/paternalism/
Fleurbaey, M. and Schokkaert, E. 2013. Behavioral welfare economics and redistribution, *American Economic Journal: Microeconomics,* vol. 5, 180–205
Friedman, M. 1966. The methodology of positive economics, pp. 3–16, 30–43 in Friedman, M. (eds) *Essays in Positive Economics*, Chicago, University of Chicago Press
Friedman, M. and Savage, L. J. 1948. Utility analysis of choices involving risk, *Journal of Political Economy,* vol. 56, no. 4, 279–304
Gigerenzer, G. 2015. On the supposed evidence for libertarian paternalism, *Review of Philosophy and Psychology*, vol. 6, 361–383, esp. 362–363
Guala, F. and Mittone, L. 2015. A political justification of nudging, *Review of Philosophy and Psychology,* vol. 6, 385–95
Hausman, D. M. and Welch, B. 2010. Debate: to nudge or not to nudge, *Journal of Political Philosophy,* vol. 18, 123–136
Kahneman, D. and Tversky, A. 1979. Prospect theory: an analysis of decision under risk, *Econometrica,* vol. 47, 263–91
Kapsner, A. and Sandfuchs, B. 2015. Nudging as a threat to privacy, *Review of Philosophy and Psychology,* vol. 6, 455–68
Lecouteux, G. 2015. In search of lost nudges, *Review of Philosophy and Psychology,* vol. 6, 397–408
Manne, G. A. and Zywicki, T. J. 2014. Uncertainty, evolution, and behavioral economic theory, *Jounrnal of Law, Economics and Policy*, vol. 10, no. 3, 555–80
Manzini, P., Mariotti, M. and Tyson, C. 2011. 'Manipulation of Choice Behavior', Working Paper No. 5891, Institute for the Study of Labor (IZA)
McQuillin, B. and Sugden, R., 2012. Reconciling normative and behavioural economics: the problems to be solved, *Social Choice and Welfare,* vol, 38, 553–567
Parfit, D. 1986. *Reasons and Persons*, Oxford Paperbacks. Oxford University Press, New York, NY
Rebonato, R. 2012. *Taking Liberties: A Critical Examination of Libertarian Paternalism, New York,* Palgrave
Rizzo, M. J. and Whitman, D. G. 2009A. Little brother is watching you: new paternalism on the slippery slopes, *Arizona Law Review,* vol. 51, 740
Rizzo, M. J. and Whitman, D. G., 2009B. The knowledge problem of new paternalism, *Brigham Young University Law Review,* 2009, 905–968
Rubinstein, A. and Arad, A. 2017. 'The People's Perspective on Libertarian-Paternalistic Policies', Working Paper, unpublished manuscript
Rubinstein, A. and Salant, Y. 2012. Eliciting welfare preferences from behavioural data sets, *Review of Economic Studies,* vol. 79, 375–387
Selinger, E. and Whyte, K. 2011. Is there a right way to nudge? The practice and ethics of choice architecture, *Sociology Compass,* vol. 5, 923–935
Sudgen, R. 2017. Do people really want to be nudged towards healthy lifestyles? *International Review of Economics,* vol. 64, 113–123
Sudgen, R. 2018. 'Better off, as judged by themselves': a reply to Cass Sunstein, *International Review of Economics,* vol. 65, 9–13
Sunstein, C. 2015A. The ethics of nudging, *Yale Journal on Regulation,* vol. 32, 414–450
Sunstein, C. 2015B. Nudges, agency, navigability, and abstraction: a reply to critics, *Review of Philosophy and Psychology,* vol. 6, 511
Sunstein, C. R. 2018. 'Better off, as judged by themselves': a comment on evaluating nudges, *International Review of Economics,* vol. 65, 1–8
Sunstein, C. R. and Thaler, R. H. 2003. Libertarian paternalism is not an Oxymoron, *University of Chicago Law Review,* vol. 70, 1159–1202
Thaler, R. H. 2015. *Misbehaving: the Making of Behavioral Economics*, New York, W. W. Norton & Company

Thaler, R. H. and Benartzi, S. 2004. Save more tomorrowtm: using behavioral economics to increase employee saving, *Journal of Political Economy,* vol. 112, S164–S187

Thaler, R. H. and Sunstein, C. R. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Revised and Expanded edition, New York, Penguin Books

Velleman, J. D. 2008. The identity problem, *Philosophy and Public Affairs,* vol. 36, 221–244

Wright, J. D. and Ginsburg, D. H. 2012. Bethavioral law and economics: its origins, fatal flaws, and implications for liberty, *Northwestern University Law Review,* vol. 106, 1090

Zywicki, J. 2017. 'Do Americans Really Save Too Little and Should We Nudge Them to Save More? The Ethics of Nudging Retirement Savings. George Mason Law & Economics', Research Paper No. 17-03, available at https://ssrn.com/abstract=2901173 or http://dx.doi.org/10.2139/ssrn.2901173