Behavioural Public Policy (2022), 1–12
doi:10.1017/bpp.2022.19

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Banishing the inner Econ and justifying paternalistic nudges

Daniel M. Hausman

Center for Population-Level Bioethics, Rutgers University, New Brunswick, NJ, USA
Correspondence to: E-mail: dhausman@cplb.rutgers.edu

(Received 16 February 2022; revised 13 May 2022; accepted 6 June 2022)

Abstract

Paternalistic nudging and framing aim to correct flaws in deliberation by relying on the same cognitive mechanisms that create those flaws. Regarding some choices as flawed and in need of correction requires some standard of correctness. In their well-known book, *Nudge*, Thaler and Sunstein take the individual's own "purified" preferences to be that standard, which is inconsistent with the finding of behavioral economics that individuals do not have a stable preference ranking of alternatives, but instead construct their preferences when faced with a choice. This essay defends an alternative, readily usable standard to judge whether individuals are choosing badly and whether nudges can help them to choose better.

Keywords: nudging; framing; preferences; well-being; paternalism

There has been a surge of empirical investigations by economists and psychologists who have identified systematic ways in which choice behavior diverges from models of self-interested rational choice. Rather than carrying around with them an all-purpose complete and transitive preference ranking, people tend to generate, on the fly, their preferences among the immediate objects of choice (Lichtenstein & Slovic, 2006). Of course, there are constraints. Maryanne, who has been a vegetarian for years, is unlikely suddenly to order a sirloin steak from the menu rather than stuffed squash. But among the unfamiliar vegetarian entrees at a Chinese restaurant, she may make up her rankings, which might have been different at another time of the day or in different company, or even if the menu had been printed in a different font. In a different context or with a different prompt, many of the preferences individuals express or reveal in their behavior will differ, and even within a single context, their preferences may be gappy and intransitive. Choice behavior is sensitive to a wide variety of contextual factors, psychological foibles and heuristics that lead individuals to act in ways that they themselves may regard (not necessarily correctly) as mistakes. Whether one enrolls in a retirement plan should not depend on whether it is the default option in the firm that employs you. But it does. How often customers choose

a given dish in a cafeteria should not depend on how prominently displayed the dish is. But it does. The value that an individual assigns to a coffee mug should not double when the mug is given to that individual, but it does. The list of foibles goes on and on. When I think back to the occasions where I have eaten too much or have commented too harshly on a student's essay, I recognize myself in the portraits behavioral economists paint of fallible human choosers.

Although it takes no great acumen to recognize that individuals sometimes make imprudent choices, economists had assumed that mistaken choices were unsystematic and could be treated as mere noise. But with the identification of systematic violations of economic rationality, economists could no longer suppose that, as a good first approximation, individuals choose whatever is best for themselves. The argument that paternalism is impossible (because individuals always choose what is best for themselves) collapses, and paternalistic policies need to be assessed on their merits. One attractive possibility is to seek out policies that will help individuals to avoid making bad choices without coercing them, by making use of the same flaws in deliberation that lead to mistaken choices.

Such policies are intriguing, but also puzzling. By what standards can policymakers judge that voluntary choices are *mistakes*? How can government help individuals make better decisions about how much to save, what to eat, whether to be vaccinated or what cars to drive? This essay defends a simple answer: there are prudential rules of thumb about what constitute good choices that apply broadly, though not universally. Policies that avoid coercion while making it easier for people to make the choices that these generalizations favor can make people better off. For example, even though there is no way for policymakers to know whether displaying healthier foods in a cafeteria more prominently benefits any particular individual, displaying healthier foods more prominently will do little harm and is likely to benefit many people.

The next section, "Nudging," clarifies the so-called libertarian paternalist proposal, which aims to improve choices by changing how choices are structured. The following section, "Conceptualizing mistakes," discusses a challenge to the whole idea of "improving" people's choices: how can economists conceive of choices as mistaken without abandoning the view that choices are a guide to well-being and imposing their values on the populace? This section presents a response to this challenge that purports to rely on the individual's own "true" preferences as the standard to judge whether choices are mistaken and to justify paternalist interventions. The penultimate section, "Reconceiving paternalistic nudging," criticizes this account of how to judge whether choices are mistaken, and it defends the alternative justification for paternalist policies sketched above. The essay concludes with the section, "Conclusion: away with the ghostly nudge judge."

Nudging

In their widely cited book, *Nudge* (2008), Thaler and Sunstein discuss the possibilities of influencing people's choices by restructuring the "architecture" of their choices. For example, exit gates at parking ramps require that customers retrieve their credit cards before the exit gates open. That design makes it much less likely that careless drivers

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

will leave their cards behind. Through "nudges" such as this one, institutions, including governments, can get people to make better choices *without limiting their freedom*.

Thaler and Sunstein define a nudge as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (Thaler & Sunstein, 2008:6). Sunstein expands this characterization of nudges as follows:

> Nudges are interventions that steer people in particular directions but that also allow them to go their own way. A reminder is a nudge; so is a warning. A GPS nudges; a default rule nudges. To qualify as a nudge, an intervention must not impose significant material incentives (including disincentives). A subsidy is not a nudge; a tax is not a nudge; a fine or a jail sentence is not a nudge. To count as such, a nudge must fully preserve freedom of choice … Some nudges work because they inform people; other nudges work because they make certain choices easier; still other nudges work because of the power of inertia and procrastination (Sunstein, 2015: 417).

As Thaler and Sunstein define nudges, they consist of any way of influencing choices that preserves freedom of choice. So defined, nudges could include, at one extreme, slipping some valium into one's rich and stingy uncle's dinner before asking for a loan, or, at the opposite extreme, offering a group of voters a course in decision theory. Actions that aim to shape choices non-coercively are heterogenous, and it is questionable whether much can be said about them all, other than that they are not coercive and they aim to influence behavior.

Methods of influencing behavior obviously differ, and they can be employed toward admirable or despicable ends. Nudges need not be employed paternalistically. They are devices for influencing behavior without coercion, regardless of the purposes. Marketing is the home of nudging. However, in this essay, I will be concerned exclusively with paternalistic nudges. Rather than using "nudging" to refer both to non-coercive persuasion in general and to the use of choice architecture, I call non-coercive influencing of any sort "steering", to avoid conflating it with nudging in the narrow sense, which I take to be a method of steering people by structuring the architecture of choice circumstances.

Paternalistic nudging in the narrow sense consists in "choice architecture" – structuring the choice situation to help individuals to make better choices non-coercively by making prudent choices easier to make than imprudent choices. The development and examination of this method of steering choices derive from the insights of behavioral economics, unlike other methods of non-coercive influence, such as "boosting" (Grüne-Yanoff & Hertwig, 2016), inflaming, flattering, socially influencing, educating, rationally persuading or confusing people. All of these fall under the expansive rubric of what Thaler and Sunstein call "nudging," but, unlike nudging in the narrow sense, they have been well known for millennia.

I focus on nudging conceived of as influencing choices by modifying the circumstances of choice, because this method fits the central examples that Thaler and

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

Sunstein give, and it is only in this mechanism of steering choices that there is any novelty in Thaler and Sunstein's work.[1] Changing the default on a retirement plan, so that employees are automatically enrolled unless they opt out is not informing, educating, flattering, confusing or inciting them. What Thaler and Sunstein' are concerned with, and what makes their work original and important is the recognition that the heuristics and flaws that behavioral economics have identified in human decision-making constitute not only problems for public policies, but also opportunities for solving them.[2]

Conceptualizing mistakes

To speak of choices as *mistaken*, rather than merely different or varied, requires some standard for choosing correctly. In some cases, this is easy. In *Hamlet*, Gertrude poisons herself by drinking wine that, unbeknownst to her, had been poisoned. Although she preferred to drink some wine over abstaining, she did not prefer to poison herself. In terms of her own preferences, once her beliefs have been corrected, her drinking the poison was a mistake. Mistakes owing to false beliefs or ignorance are easy to explain.

It is not so easy to determine when choices that are influenced by rational foibles rather than false beliefs constitute mistakes. Consider what might appear to be an easy case of weakness of will. During our Wednesday morning walk, my friend, Mary, shamefacedly confesses that at a dinner party last night, she not only had chocolate mousse for dessert, but she then helped herself to a second portion. "I don't know what got into me. It was so delicious. But I wish I had more self-control." Knowing that she is a borderline diabetic, I judge that she is right to regret her choice. She placed herself at risk of a diabetic coma, aggravated her high cholesterol and failed to adhere to the diet to which she was committed. She also feels sluggish and headachy this morning, which is how she usually feels after eating sweets. Mary and I both judge that dessert last night was a mistake.

Was it? That's how it looks on Wednesday morning. But on Tuesday evening, the mousse was there in all its chocolate splendor, and when the mousse was passing around the table, the risks and harms of consuming it were less salient to Mary than the pleasures that beckoned. Her choice was not due to ignorance. She knew that the dessert was chocolate mousse that it contained scrumptious quantities of sugar and fats and that eating the mousse involved a significant risk to her health and a violation of her diet. Yet she helped herself to the mousse. No one else moved her hand.

At that time, did she *prefer* to eat the mousse? If one adheres to the modeling strategy that economists employ, whereby everything that determines choices, apart from

[1]"What is genuinely novel about the nudging approach … is the idea of exploiting people's cognitive and motivational deficiencies in ways that help them to make decisions that their better self (or superego) would make" (Grüne-Yanoff & Hertwig, 2016: 153).

[2]Here I follow Congiu and Moscati, "we conceive of nudges as any attempt at influencing people's behaviour that instrumentally exploits their rationality failures, such as their cognitive boundaries, biases and habits (2020: 71–72) and Mongin and Cozic (2018: 2) who define what they call 'Nudge 2' as 'an intervention that uses rationality failures instrumentally'."

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

beliefs and physical constraints, does so via influencing preferences, then the answer must be "yes." There was no constraint, nor was there any false belief. As much as she now regrets her choice, it was her choice, determined by her preferences. From her perspective (at least on Wednesday morning), the sight and smell of the mousse "distorted" her preferences, but distorted or not, on Tuesday night she preferred to indulge.

Yet she would repudiate those preferences now. Does that make them *mistaken*, rather than merely different? Some mornings I prefer toast to cereal for breakfast. Other mornings I feel like cereal. Yet I don't think that I make a mistake every time my preference changes. Nor does Mary think that she makes a mistake when, on other occasions, she shows more will power and refuses dessert. "But we still need a criterion of individual welfare or well-being to substitute for the traditional criterion of preference satisfaction. Without such a criterion, we cannot know in which directions individuals should be nudged" (McQuillin & Sugden, 2012: 560).

One way of making sense of the claim that the mousse indulgence was a mistake is to maintain that Mary identifies with one of her preference orderings. She believes, both on Tuesday night and on Wednesday morning that foregoing the mousse is the more choiceworthy alternative for her. The preference ordering that places abstaining over indulging is, in her view, the more authentic ranking, the ranking that expresses her real values, who she really is. (I do not assume that she is right about this. It may be that she is in truth a vacillating person who frequently indulges, albeit with guilt and regret, or it may be that she is more accurately portrayed as a repressed epicure who has not yet shaken off an ascetic upbringing.) Choices that violate the ordering of alternatives that conforms to "what truly matters" to Mary are mistakes.

As Infante *et al.* (2016a, 2016b) (ILS) and Hands (2020) point out, conceiving of mistakes this way is a perilous endeavor. Thaler and Sunstein write, "So long as people are not choosing perfectly, some changes in the choice architecture could make their lives go better (*as judged by their own preferences*)" (2008: 10). Thaler and Sunstein declare emphatically that their aim is "to make people better off *as judged by themselves*" (2008: 5). The problem with this is that if choices are determined by preferences, then people's choices before they are nudged are already the best they could be "as judged by their own preferences." The only way out of this impasse that does not surrender the connection between preference and well-being is to maintain that what is chosen is suboptimal with respect to some other preference ranking that differs from the individual's manifest preferences. The idea is to treat "cases in which an individual's choices depend on 'irrelevant' properties or framing as errors, 'error' being defined relative to the preferences that the individual would have revealed if not subject to reasoning imperfections" (ILS, 2016a: 9).

Infante *et al.* describe this move as in effect, hypothesizing that Humans, who are subject to these rational foibles and, especially, context-dependent preferences, possess within them an "inner rational agent," whose preferences, like those of "Econs," are context-independent and free of the distortions that characterize Humans. This inner agent is "the locus of the identity of the human being and … the source of normative authority about its interests and goals" (2016a: 1).

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

> [B]ehavioural welfare economics models human beings as faulty Econs. Its implicit model of human decision-making is that of a neoclassically rational inner agent, trapped inside and constrained by an outer psychological shell. Normative analysis is understood as an attempt to reconstruct and respect the preferences of the imagined inner Econ (ILS, 2016a: 22).

As many have noticed (McQuillin & Sugden, 2012; ILS, 2016a; Dold & Schubert, 2018; Sugden, 2018b; Hands, 2020), mistakes in this sense are of questionable relevance to policymakers who are not in a position to know what Mary's authentic preferences are and who would seem to be overstepping the bounds of liberal state policy in aiming to prevent members of the populace from betraying their true values. When mistakes are conceived of in this way, policies devoted to helping people to avoid mistakes must attribute to individuals a preference ordering that captures their real values, which may conflict with the preferences that are manifest in the choices they make. Identifying "true" preferences has been described in the literature as "preference purification," but extracting true preferences from choices and verbal behavior purifies preferences in different ways. In Mary's case, it involves repudiating preferences that are allegedly not in accord with Mary's values. Policymakers would be on shaky ground in attempting to help individuals to avoid such mistakes.

Moreover (and more seriously) postulating purified preferences as the standard in terms of which to judge whether people's choices are mistaken apparently jettisons one of the most important findings of behavioral economics, which is that rather than carrying around a complete and transitive ranking of all possible alternatives, people typically construct their context-dependent preferences in the course of making a choice (Lichtenstein & Slovic, 2006). To suppose that there is a "purified" context-independent preference lurking within is fundamentally at odds with the picture of individuals constructing their preferences when they are called for (Hands, 2020). If one takes the findings of behavioral economics seriously, it is a mistake to speak of "purifying" context-dependent "distorted" preferences.[3] Instead of tidying up an existing preference ranking, economists are inventing a preference ordering where none exists.

In the case of preferences that are distorted by ignorance or false beliefs or that are not entirely self-interested, one can speak sensibly of "cleaning up" an existing preference ordering. In Gertrude's case, if her beliefs had been true, there would have been nothing wrong with her choice. Unlike Mary's indulgence, there is no question here concerning which are Gertrude's true preferences.[4] When the actions of the populace are governed by false beliefs, like so much of the resistance to vaccination against COVID-19, then there is room for policy measures to persuade or coerce individuals to take actions that benefit themselves and others from the perspective of their own preferences, as they would be if their beliefs were not erroneous. But when the problems lie in alleged psychological distortions within preferences, as in Mary's case, policies devoted to "improving" choices rest on presumptuous and hard-to-justify

[3]See also the criticisms by Dold and Schubert (2018: 280).

[4]Actually, there is a rather different question one might ask: Does she know that the wine is poisoned, and drinks it to protect Hamlet?

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

assumptions concerning individual's hidden preferences, whose existence behavioral economics gives us good reasons to doubt.

This line of thought apparently leads to the conclusion that nudging is irretrievably confused. But something has surely gone wrong in this argument. There's good reason to encourage firms to make contribution to retirement plans the default for their employees and to encourage cafeteria managers to display healthier foods more prominently and attractively. How can we make sense of the considerations in favor of paternalistic nudging?

Reconceiving paternalistic nudging

The view I shall sketch on behalf of Camerer *et al.* (2003), Sunstein and Thaler (2003) and Thaler and Sunstein (2003, 2008) is not faithful to what they write, but it leads to a coherent account of nudging that explains how it can be useful and consistent with the findings of behavioral economics.

Consider what is probably the most successful example of nudging: setting the default in retirement plans so that employees are automatically enrolled, although they can easily opt out if they prefer. When faced with choices among retirement plans, presumably some individuals have stable preferences, and their choice of whether to enroll or not is not influenced by whether enrollment is the default. Those who have such fixed preferences concerning retirement do not necessarily have stable preferences between abstaining from and indulging in multiple portions of chocolate mousse. They are not Econs, even if they have firm and well thought out preferences among retirement options.

What makes nudging employees to sign up for retirement plans potent is the apparent fact that many individuals do not have fixed preferences among retirement plans and instead construct their preferences among them "on the fly." If the default is not to enroll, many people, through laziness or some heuristic or other, do not enroll, even though with a different default they would have constructed other preferences. When asked, they will probably offer rationalizations for their choices. Regarding nudges as influencing preference formation rather than as enabling hidden preferences to show themselves is more plausible than is the story of Econs and Humans and in better accord with the findings of behavioral economics.

If nudging influences preference formation, then it may be the case that people are "better off *as judged by themselves*," whether nudged to save for retirement or whether nudged not to save for retirement. What sense can we make of the view that nudges are successful if they "make people better off *as judged by themselves*" (Thaler & Sunstein, 2008: 8), when the preferences that ground those judgments may be made up on the spot?

First, note an ambiguity. Is what is at issue preferences among nudges or preferences among the outcomes toward which one is nudged? In contrast to Cartwright and Hight (2020), I interpret the object of the judgment to be the outcome of the nudge, not the experience of having been nudged, which individuals may not even be aware of. Second, does the requirement that people be better off "as judged by themselves" demand both that individuals are, in fact, better off and that they judge that they are, or does it demand only that individuals believe that they are

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

better off? When Thaler and Sunstein first introduce the objective of making individuals better off "as judged by themselves," they write, "we argue for self-conscious efforts …to steer people's choices in directions that will improve their lives" (2008: 5). They are apparently concerned with improving individual's lives, which they assume individuals will recognize, not merely with making people believe that their lives are better. Yet, the first interpretation, which demands that nudging make people genuinely better off, seems to be ruled out by Thaler and Sunstein's insistence that they are not committed to any view of which is the better choice, full stop. They insist that they have no interest in telling people how to behave. When they assert that "some changes in the choice architecture could make their lives go better" they immediately add, "(as judged by their own preferences, not those of some bureaucrat)" (pp. 9–10). They claim to rely entirely on the values of the agents themselves. Hence, all they can mean when they say that paternalistic nudges "make people better off *as judged by themselves*" is that people believe that successful nudges make them better off.

If, as I have argued, choice architecture influences preferences and not just how preferences display themselves in choices, then individuals may be just as satisfied with their choices regardless of which way they are nudged. In that case, why believe that nudging makes people better off? Those who only care about whether people judge that they are better off might not find this possibility unnerving. All that matters is whether individuals are satisfied with their choices. There is no reason to invoke an inner rational agent, unless one wants to make people better off and needs the judgments of an inner rational agent to provide a criterion that goes beyond contentment.

Those who enroll in a retirement plan because enrollment is the default are unlikely to have any settled view about how much better off they are than if they had not enrolled – probably not for decades in any case (Cartwright & Hight, 2020). If they do think they are better off, it may be due to cognitive dissonance, which would have led them to be just as satisfied with the opposite choice. To the extent that the question ever occurs to Nancy and George (who ate at the cafeteria last week) whether the prominently displayed fruit they chose made them better off than the cake they usually take, they are bound to have a variety of views. How large a consensus is needed? Does the justifiability of the nudge depend on how many are happy with their choices?

The way out of this morass and at the same time to respond to ILS's and Hands' criticisms of the postulation of an inner rational agent is to give up the hope that libertarian paternalism can function without any normative premises other than that it is good if people are better off "as judged by themselves." It is good if people are better off "as judged by themselves" provides a criterion by which to judge the success of nudging only if being better off "as judged by themselves" has some relationship to being better off, full stop. Thaler and Sunstein clearly assume that there is some such relationship, since they aim to "steer people's choices in directions that will improve their lives" (2008: 5).

In recent work, Sunstein attempts a response to this criticism. He writes, "Many nudges are designed to make people better off, as judged by themselves. This criterion, *meant to ensure that nudges will increase people's welfare*, contains some ambiguity." (2018: 1 [italics added]). If one assumes that nudges help people to satisfy fixed

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

antecedent "purified" preferences, and those who are nudged evaluate their actions by the extent to which they satisfy those preferences, then making people better off as judged by themselves is making people better off. Given these assumptions, nudges increase welfare whenever people believe that they are better off. However, prompted by Sugden's criticisms (2017), Sunstein recognizes that it may be the case that, nudged toward *x*, Nancy would judge *x* as better for her than *y*, while nudged toward *y*, she would judge *y* as better for her than *x*. In such a case, her judgment does not determine which enhances her welfare.

> Sunstein responds
> Social planners – or in our terminology, choice architects – might well have their own ideas about what would make choosers better off, but in our view, the lodestar is people's own judgments. To be a bit more specific: 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: 2).

The first sentence sounds as if it reaffirms the insistence on deferring to the judgments of those who are nudged, but the second sentence (which is hardly just "a bit more specific") takes that back. It implies that the objective is to make people better off, *regardless of their judgment*, which is only "a good (if sometimes imperfect) way to test the question whether nudges are increasing their welfare." Sunstein is saying that the views of the nudged are only evidence concerning what increases people's welfare. The approval of the nudged is no proof of an improvement in welfare. This is, in my view, unobjectionable; however, uncomfortable it may make Thaler and Sunstein and other economists who would prefer to delegate normative judgments to someone else.[5]

Recognizing that approval of the nudged is not decisive means that Thaler and Sunstein should concede that they rely on substantive premises concerning which choices generally make people better off. Policymakers are in no position to know whether Nancy is better off saving for retirement, or whether George will wind up unable to pursue some important expensive goal because of mistakenly enrolling in a retirement plan. What policymakers can judge – fallibly, of course – is that a double portion of chocolate mousse is worse for most middle-aged overfed Americans than one or none or that Americans generally save too little for retirement.[6] Such generalizations, which typically have not been rigorously tested, are a questionable basis for coercive policies, but they are arguably sufficient to justify the placing the fruit in a more prominent place and setting contribution as the default in a retirement plan, since it is easy for those who want the cake rather than the fruit or present consumption rather than saving for retirement to do as they please.

[5]Robert Sugden attempts to reconcile the sentences in this quotation by arguing that the evidential connection between people's judgments and what is good for them that the second sentence asserts justifies the deference to people's own judgments the first sentence urges (2018a: 10).

[6]One sees such judgments in *Nudge* itself, despite Thaler and Sunstein's claim to defer to the judgments of the nudged. "With respect to diet, smoking, and drinking, people's current choices cannot reasonably be claimed to be the best means of promoting their well-being." (Thaler & Sunstein, 2008: 7).

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

Generalizations concerning which choices generally make people better or worse off make it possible to judge whether nudging benefits individuals. There is no need to invoke some hidden preference ranking that can be foisted on individuals after sufficient "purification" of their manifest preferences. Rather than justifying nudging people not to overeat by supposing that down deep overeaters possess inner Econs who prefer not to eat too much, generalizations such as "overeating is on average harmful to individuals" can justify nudging people to eat moderately. Thaler and Sunstein implicitly invoked an inner rational agent to do a specific job: to serve as a criterion for whether a nudge is beneficial. Since prudential rules of thumb can do the job, the inner rational agent can be retired (without a pension).[7]

Can these rules of thumb, which float in their own limbo, do the job? Without some defensible account of what constitutes well-being, how can these generalizations *justify* policies designed to promote well-being?[8] From the perspective of philosophical theory, they are unsatisfactory. They dodge all the hard questions concerning the composition of well-being. But it seems to me that in order to justify implementing a paternalistic nudge, policymakers (or benevolent cafeteria owners) do not need anything more than generalizations such as "fruit is a healthier desert than chocolate mousse," "It is good to be healthier," and "There is no other generally important benefit to eating mousse rather than fruit." These are not only uncontroversial; there are empirical or theoretical grounds that support them, and they face no serious criticism. If nudges can lead to more choices of fruit for dessert without coercing individuals or incurring other significant costs, policymakers who are charged with promoting well-being have good reason to nudge. This conclusion relies on the assumption (which is contestable) that the agents who are considering implementing nudges have the authority to institute non-coercive policies that promote virtually uncontested benefits. Since nudging is rarely free, nudging by the government will be supported by taxation; and it will consequently not be completely non-coercive. But this sort of coercion is involved in the provision of all social policy.

Thaler and Sunstein will not be happy with my claim that what justifies nudges are generalizations about what typically makes people better off, whether or not the agents see it that way. I am not defending this view as an interpretation of their work, but instead as a more straightforward justification for nudging than theirs. What is especially attractive about nudging is that policymakers can institute policies that they believe to be generally prudent with low risk, since nudges are weak, and agents are free to go their own way. What justifies an apparently successful nudge, such as painting "Look right" and "Look left" at street corners in London, is whether the warnings avert accidents. Of course, in this case, it is uncontroversial to maintain that with very few exceptions, people have a laundered preference not to step into the road looking the wrong way, but there is no need to invoke these preferences.[9]

[7]Unlike my argument (2016), which contests ILS's claim that Thaler and Sunstein's position commits them to an inner rational agent, this essay argues that ILS are right, but that there is an alternative way to justify nudging.

[8]I am indebted to Robert Sugden for this objection.

[9]As is the case with many paternalistic policies, there is also a non-paternalistic justification for the painted directions. The warnings avert the costs that would be imposed on Londoners by the carnage inflicted on tourists. I am indebted here to an anonymous referee.

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

Thus, there is a simple way in which one can judge Mary's mousse debauch to be a mistake: it could be expected to be harmful to her, where the standard of harm is provided by generalizations about what sorts of choices are good for people, rather than by Mary's "true" preferences. Policymakers are unlikely to know what sort of diet is best specifically for Mary, but they can know what sorts of diet are healthier for most people (in a relevant reference class), and they can use that knowledge, coupled with the assumption that better health is good for most people, to design policies to make people better off. Since there is no reason to believe that two helpings of chocolate mousse are bad for everyone, there are good reasons to prefer less coercive means of steering choices such as nudging.

This justification for nudging leaves many questions unanswered. How much evidence do policymakers need? Nudging is unlikely to be free. How does one measure the benefits of nudging and balance them against the costs? Will a better justification for nudging encourage an increasing tolerance for coercive policies (Rizzo & Whitman, 2009)? Will nudging provoke resentment and thus backfire? Will policymakers be able to identify nudges that are effective, impose no serious costs on those who resist them and have no other drawbacks?

Conclusion: away with the ghostly nudge judge

Having jettisoned the "as judged by themselves" criterion for successful nudging, there is no need to invoke Econs within to serve as the judges of nudges. We can exorcize those ghosts without, as Sugden urges (2018b), giving up a concern with well-being or with preference satisfaction as an indicator of well-being. By invoking a notion of well-being that is independent of the beliefs of the nudged, there is no need to invoke an inner rational agent to appraise nudges, and very good reason not to. The conflicts with the findings of behavioral economics thereby dissolve. Nudging uses choice architecture to make people better off. That aim is fully consistent with a recognition that people's preferences are often constructed at the moment of choosing, and that they are incomplete, context-independent and often intransitive. How much good nudging can do is uncertain, but libertarian paternalism need not share Thaler and Sunstein's implicit commitment to an inner rational agent.

Conflict of interest. There are no competing interests or conflicts of interest.

References

Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue and M. Rabin (2003), 'Regulation for conservatives: behavioral economics and the case for 'asymmetric paternalism'', *University of Pennsylvania Law Review*, **151**: 1211–1254.

Cartwright, A. and M. Hight (2020), '"Better off as judged by themselves": a critical analysis of the conceptual foundations of nudging', *Cambridge Journal of Economics*, **44**: 33–54. https://doi.org/10.1093/cje/bez012.

Congiu, L. and I. Moscati (2020), 'Message and environment: a framework for nudges and choice architecture', *Behavioural Public Policy*, **4**: 71–87.

Dold, M. and C. Schubert (2018), 'Toward a behavioral foundation of normative economics', *Review of Behavioral Economics*, **5**: 221–241.

https://doi.org/10.1017/bpp.2022.19 Published online by Cambridge University Press

Grüne-Yanoff, T. and R. Hertwig (2016), 'Nudge versus boost: how coherent are policy and theory?', *Mind and Machines*, **26**: 129–183.
Hands, D. W. (2020), 'Libertarian paternalism: taking Econs seriously', *International Review of Economics*, **67**: 419–441.
Hausman, D. (2016), 'On the Econ within', *Journal of Economic Methodology*, **23**: 26–32.
Infante, G., G. Lecouteux and R. Sugden (2016a), 'Preference purification and the inner rational agent: a critique of the conventional wisdom of behavioural welfare economics', *Journal of Economic Methodology*, **23**: 1–25.
Infante, G., G. Lecouteux and R. Sugden (2016b), 'On the Econ within': a reply to Daniel Hausman', *Journal of Economic Methodology*, **23**: 33–37.
Lichtenstein, S. and P. Slovic (2006), 'The Construction of Preference: An Overview', in S. Lichtenstein, and P. Slovic (eds), *The Construction of Preference*, New York: Cambridge University Press, 1–40.
McQuillin, B. and R. Sugden (2012), 'Reconciling normative and behavioural economics: the problems to be solved', *Social Choice and Welfare*, **38**: 553–567.
Mongin, P. and M. Cozic (2018), 'Rethinking nudge: not one but three concepts', *Behavioural Public Policy*, **2**: 107–124.
Rizzo, M. and D. Whitman (2009), 'Little brother is watching you: new paternalism on the slippery slopes', *Arizona Law Review*, **51**: 685–739.
Sugden, R. (2017), 'Do people really want to be nudged towards healthy lifestyles?', *International Review of Economics*, **64**: 113–123.
Sugden, R. (2018a), '"Better off, as judged by themselves': a reply to Cass Sunstein', *International Review of Economics*, **65**: 9–13.
Sugden, R. (2018b), *The Community of Advantage: A Behavioural Economist's Defence of the Market*. Oxford: Oxford University Press.
Sunstein, C. (2015), 'The ethics of nudging', *Yale Journal on Regulation*, **32**(2015): 413–450.
Sunstein, C. (2018), '"Better off, as judged by themselves": a comment on evaluating nudges', *International Review of Economics*, **65**: 1–8.
Sunstein, C. and R. Thaler (2003), 'Libertarian paternalism is not an Oxymoron', *University of Chicago Law Review*, **70**: 1159–1202.
Thaler, R. and C. Sunstein (2003), 'Behavioral economics, public policy, and paternalism', *American Economic Review*, **93**: 175–179.
Thaler, R. and C. Sunstein (2008), *Nudge*. New Haven: Yale University Press.

Cite this article: Hausman DM (2022). Banishing the inner Econ and justifying paternalistic nudges. *Behavioural Public Policy* 1–12. https://doi.org/10.1017/bpp.2022.19