Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

Nudges, Nudging, and Self-Guidance Under the Influence

W. JARED PARMER
RWTH Aachen University

Nudging works through dispositions to decide with specific heuristics, and has three component parts. A *nudge* is a feature of an environment that enables such a disposition; a person *is nudged* when such a disposition is triggered; and a person performs a *nudged action* when such a disposition manifests in action. This analysis clarifies an autonomy-based worry about nudging as used in public policy or for private profit: that a person's ability to reason well is undermined when she is nudged. Reasoning well is a component of self-guidance, which is an aspect of autonomy and so something there is reason to promote, preserve, or respect. However, a person can reason well when she is nudged: Many of these heuristics are good rules to reason with, and she can be flexible with respect to them when she takes there to be a better way to reason. Along the way, this paper uncovers several design specifications for responsible nudging, and discusses open empirical questions. However, nudging's being compatible with reasoning well crystallizes a distinct worry about manipulation: that nudge designers can rely on nudged people guiding themselves toward the designers' own ends. Manipulation of this sort exploits one aspect of autonomy (namely, self-guidance) to undermine autonomy in other respects.

Imagine that a surgeon wants a patient to consent to a surgery that carries some serious costs, but that has the potential to dramatically improve the patient's quality of life. Suppose, for example, that the surgeon will attempt to fully remove a cancerous growth from his throat, but might be unsuccessful; and, either way, the surgery will require the patient to use a gastrostomy tube for an extended period afterward, and to cope more generally with the pains of recovery. However, the surgeon suspects that her patient will give these latter costs too much weight and might refuse the surgery on those grounds. And she has good reason to suspect that, if she frames the potential of the surgery

Contact: W. Jared Parmer <jared.parmer@humtec.rwth-aachen.de>

https://doi.org/10.3998/ergo.2913

in the right way, she can make it more likely that he will consent. In particular, the surgeon considers saying that four out of five patients in comparable situations are cancer-free for at least five years as a result of the surgery, rather than that one out of five patients continue to suffer from the cancer after the surgery—where, in either case, she would also fully inform him about the G-Tube and other difficulties that arise as a result of the surgery. Should she opt for the former framing over the latter one? Should she *nudge* her patient to get the surgery?

Or imagine that a tech firm has hired a "Wellness Guru" to design a cafeteria at headquarters for its employees. The Wellness Guru believes that, if the employees eat a balanced and healthy diet, they will be happier and more productive at work. And he believes that he can make this more likely by designing the cafeteria to guide the employees along a certain path, with healthier items in more prominent locations along that path, and less healthy items relegated elsewhere. So, for example, he might place well-lit bowls of fruit right at the entrance and trays of cookies underneath the utensils at the end, a bit below waist-level. Alternatively, he could design the cafeteria with an entirely open floorplan, with each broad category of item clearly marked, equally visible and accessible from the main entrance. Should he opt for the former layout over the latter? Should he *nudge* the employees to eat healthier?[1]

These are examples of a phenomenon that has come under intense debate.[2] When, if ever, should we nudge one another? Clearly, this question hangs on what nudging *is*.[3]

So far, many accounts on offer characterize nudging as an ethically loaded kind of influence. Here are a few representative examples, the first from Bart Engelen and Thomas Nys:

1. For a similar example, see Thaler and Sunstein (2008: 1).

2. The literature is vast and growing at a rapid clip, but see Thaler and Sunstein (2008) for the opening salvo in favor of nudging, and Bovens (2009) and Hausman and Welch (2010) for highly influential replies. Within the last couple of years, numerous philosophers have come to the defense of nudging; see, for example, Engelen (2019), Engelen and Nys (2020), Houk (2019), Levy (2019), and Schmidt (2019). For a comprehensive review of the ethical issues involved, consult Schmidt and Engelen (2020).

3. The *efficacy* of nudging is also, of course, relevant. But since we do not yet have a good understanding of what nudging is, we are not in a position to assess the efficacy of various purported nudging strategies. An investigation conducted by the UK House of Lords found little evidence that behavioral-change interventions, including some purported examples of nudging, were effective at the population level, despite the evidence that they can be found to be effective in short-term, controlled experiments (2011). Regarding this, the order of candidates placed on electoral ballots has a well-documented, small, but population-level effect. Ordering strategies are plausibly a way of nudging. See the discussion of ballot order effects in Section 3.1.

> [W]e take nudges to be deliberate changes in people's choice architectures with the intention of predictably influencing their behavior by tapping into a-rational psychological mechanisms—often labelled 'heuristics and biases'—and thus without merely informing, rationally persuading, incentivizing or coercing them. (Engelen & Nys 2020: 138; see also Engelen 2019: 205)

Similarly, Grant Rozeboom says that "Nudges [are] devices of 'choice architecture' that improve our choices by exploiting our mental heuristics and decision-making biases" (2020: 107). And, according to Till Grüne-Yanoff and Ralph Hertwig, nudging is distinctive because it involves "exploiting people's cognitive and motivational deficiencies in ways that help them to make decisions that their better self . . . would make" (2016: 153).

Notice that it is hard to hear these characterizations without the ring of the normative: There is evidently something pro tanto wrong with tapping into or exploiting people's non-rational biases or deficiencies to get them to do something.

Now, "nudging" and its cognates are relatively novel terms meant to capture a distinctive kind of influence that largely escaped theorists' notice, before recent deliverances of behavioral psychology and economics made us look again. The problem with the above characterizations is that they define "nudging" in terms of psychological processes with a certain (less than impressive) rational status; but, as we will see, the rational status of the processes through which nudging works is partly what is at issue in the ethics of nudging. In order words, the meaning on offer is freighted with normative baggage. So, we should want to know what nudging is, as free of normative baggage as possible, before we assess whether to do it.[4]

In Section 1, I will motivate this point further and then present my own account that builds on, but goes beyond, Yashar Saghai's (2013a; 2013b). To spoil the story just a bit, I will suggest that nudging should be understood in relation to processes that realize *dispositions to decide by using a heuristic*. This will give us a

4. Very briefly, this is why it is premature to draw a distinction between "nudging" and "boosting", as some theorists want to—most prominently, Grüne-Yanoff and Hertwig (see, e.g., Grüne-Yanoff 2018; Grüne-Yanoff & Hertwig 2016; and Hertwig & Grüne-Yanoff 2017). It might turn out that "nudging" and "boosting" refer to the very same sort of influence, construed without normative baggage. Indeed, it is striking that proponents of this distinction are trying to sketch out two very different policy programs that are in broad agreement about the psychological processes such policies are designed to work through; their disagreements begin, but do not end, with the *rational status* of these processes operating as they do in various environments (cf. Grüne-Yanoff & Hertwig 2016; for a similar point, see Engelen 2019: 221). So even if this is broadly on the right track, we would still need a neutral characterization of those processes themselves. My thanks to an anonymous reviewer for pressing me to address this.

more granular picture of these underlying processes than we have enjoyed so far, and reveals that nudging, in turn, can be decomposed into three parts: the nudge itself, an agent's being nudged, and an agent's performing a nudged action.

This account also allows us to state more precisely an ethical worry about nudging that has to do with *autonomy*. Very briefly, the worry is that, when and because an agent is nudged, she will fail to reason well about what to do or believe. If this is right, the worry continues, nudging undermines her self-guidance. I discuss and motivate this worry in Section 2. In Section 3, I reply to this worry by arguing that being nudged is compatible with reasoning well about what to do. Since this is so, nudging need not undermine self-guidance, for all that has been said. This reply also provides us with a few design specifications for responsible nudging. On both fronts, the agent's normative point of view is going to matter a great deal.

However, nudging is not in the clear as far as autonomy is concerned. In Section 4, I argue that, not only can nudges be designed to be compatible with reasoning well, but they can be designed to *rely on* it. Indeed, something stronger is true: Nudges can be designed to *exploit self-guidance* in an objectionably manipulative way. Now, suppose that autonomy is a rich and multifaceted phenomenon, of which self-guidance is only one aspect; and that, when manipulation is wrong, it is so at least partly because it undermines autonomy. Under these suppositions, nudging looks to be a distinctively pernicious threat to autonomy after all: It can be used to exploit one aspect of autonomy to undermine autonomy in other ways. My account brings this threat into sharp focus.

The principal, novel takeaways of this paper are two. First, nudging can be fruitfully accounted for in terms of our dispositions to decide by using heuristics, rather than in terms of processes with a particular rational (or nonrational) status per se. Second, a serious and underappreciated threat is that nudging might be a tool for manipulating us via our own self-guidance—if you like, partly *because* our decision-making processes are rational on some construals. This threat has received too little attention, I speculate, because we are too quick to distinguish decision-making processes in terms of their rational (or nonrational) status when we are thinking about the ethics of nudging.[5] Developing strategies to counter or mitigate this threat ought to be a focus of future research.

5. Consider that researchers working on the ethics of nudging often *define* manipulation as influence that bypasses or impedes rational decision-making processes (e.g., Blumenthal-Barby 2012; and Blumenthal-Barby & Burroughs 2012). And, in work on manipulation more generally, the non-rational status of the processes through which nudging is supposed to influence us largely goes unquestioned (as in, e.g., Gorin 2014b: 95).

1. Nudging Has Three Parts: The Nudge, Being Nudged, and Nudged Action

1.1. Motivating the Analysis

Let me draw your attention to the fact that, as we have seen, nudging is distinct from persuading, (substantially) incentivizing, or coercing, according to Engelen and Nys (2020: 138). Similarly, Andreas Schmidt says that "instead of changing people's set of options, or significantly altering their economic incentives, public policy nudges improve people's decisions by changing how options are presented to them" (2019: 512). And, lastly, Richard Thaler and Cass Sunstein say that a nudge "alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives" (2008: 6).

These characterizations are broadly on the right track with respect to the sorts of examples with which we began. The surgeon is not interested in forcing the patient to agree to the surgery; nor is she offering him additional incentives to do so. Now, she is considering *how* to present the potential and risks associated with the surgery to him, but *whether* to do so is not at issue; so the strategy she is considering is not simply about informing him of relevant facts, nor to persuade him thereby. Similarly, the "Wellness Guru" is not considering whether to force the employees to eat healthier meals, nor is he trying to substantially incentivize doing so. And he is not considering whether to straightforwardly inform or try to persuade them to eat healthier: He is not considering, for example, whether to provide nutritional information or placards that provide arguments for eating fruit, rather than cakes or cookies, for dessert.

Notice that a rough *job description* of nudging is coming into view that has the promise of being free of normative baggage, as we should want: These characterizations are in part saying what kind of job nudges perform on people who are nudged. In particular, they say what does *not* fall under nudges' remit: Nudges do not mandate or (substantially) incentivize particular decisions, nor do they straightforwardly provide information that is relevant to the decision.

Just as a theorist can give a job description of greater or lesser granularity, these functional characterizations of nudging succeed, to differing degrees, to tell us what nudging is. This is not in itself a mark against these views; it depends on what the theorists want to *do* with them. As we will see, however, one of the things we should want to do is assess the rational status of the psychological processes through which nudging works. And for this aim, we will need a bit more detail than we have got: We will need to know more about these processes themselves.

Let me show this with a brief discussion of Neil Levy's (2019) account of nudging. Levy argues that the psychological mechanisms through which nudging influences us *are* "reasoning mechanisms", and so nudges are "addressed to" reasoning mechanisms. I will not rehearse the full sweep of his argument here.[6] However, a key claim in his case is that nudges' job involves "making considerations salient to us, and disposing us to respond to them in a way that reflects their actual reason-giving force" (2019: 289).

Levy's arguments for this key claim call for a more subtle assessment than I can provide here, but let me just mark, in two steps, why this claim matters for present purposes. First, his argument hangs on what it is for a form of influence to "dispose us to respond" to a consideration "in a way that reflects its actual reason-giving force". The rub is that influences can make considerations salient to us that actually *are not* good reasons to do or believe whatever it is we are being influenced to do or believe. For example, a tobacco company's ad campaign can make it salient that some attractive people smoke, but this is not a good reason to buy their product. So Levy relies on the claim that nudging only *disposes* us to respond in the appropriate way, via some particular psychological mechanism. He then explains such faulty cases as those in which the implicated mechanisms are in some way "misfiring" (e.g., at 2019: 290), or are failing to *play their role* as reasoning mechanisms while still *having* that role (e.g., at 2019: 292).

Now, the plausibility of these sorts of moves hangs on further details about these mechanisms—details that illuminate, in a non-question-begging way, what it would be for them to "misfire" or "fail to play their role". So, second, Levy relies on the familiar theoretical apparatus of "Dual Process Theory" to put some color on our picture of these mechanisms, and grants that nudging works on us via type one processes, which are characterized in contrast to type two processes as fast, effortless, typically unconscious, and inflexible (see Kahneman 2011, among many others).[7] And this is meant to help us see what the relevant psychological mechanisms look like. Here now is the real rub: This characterization of type one processes displays, at best, *symptoms* of the underlying mechanisms at work; it does not say anything about their *nature*. Until we have that, we are not in a position to evaluate many of the key moves of his argument.[8]

---

6. Just to complete the sketch, though: Levy thinks that reasoning *just is* responding to salient considerations in the appropriate way, such that any psychological mechanism whose function is to do *that* will thereby *be* a 'reasoning mechanism' (2019: 283); that the psychological mechanisms through which nudging influences us have that very function; and that, therefore, nudging is addressed to reasoning mechanisms (for the latter two claims and their defense, see 2019: §3). In saying that nudging is 'addressed to' reasoning mechanisms, he seems to mean that nudging is *offering (putative) normative reasons*.

7. Schmidt makes a similar appeal (see 2019: 513).

8. An analogy might help here. Imagine that you were reading a paper that argued that a particular human neurological mechanism M in fact has the functional role that is constitutive of

To be clear, I am sympathetic to the drift of Levy's argument. This is not meant to be a refutation. Rather, this brief tour shows that we need to know more about the psychological processes through which nudging works if we want to do what Levy wants to: if, that is to say, we want to roll up our sleeves and assess the rational status of these processes themselves.

1.2. Building Blocks of the Analysis: Heuristic Dispositions

Let me build on Yashar Saghai's account of nudging to get us where we need to go. According to Saghai,

> A nudges B when A makes it more likely that B will x, *primarily by triggering B's shallow cognitive processes*, while A's influence preserves B's choice-set and is substantially noncontrolling. (2013a: 491, some variables changed and emphasis added; cf. Thaler & Sunstein 2008: 6)

An influence is "substantially noncontrolling", Saghai goes on to say, "if B can easily resist A's attempt to get her to x" (2013b: 499, variables changed).

Saghai says that shallow cognitive processes are marked by their speed, cheapness (in terms of effort), and the fact that they do not involve full-blown deliberation (Saghai 2013a: 489). This picture will have to be enriched, so let me suggest the following. These processes are *cognitive* because they involve estimating probabilities, diagnosing situations, drawing inferences, and making choices about what to do. They are *shallow* because these estimations, diagnoses, inferences, and choices are done on the basis of fewer considerations than are available to the person. They thus follow patterns that diverge in characteristic ways from the exhaustive, careful thinking we associate with reflection.

To illustrate with just one of the above examples, we can hypothesize that the tech employees will choose the healthier options when and because they are *satisficing*, which is to say that they choose the first option they encounter that meets some minimally acceptable threshold for what matters to them in the choice. While chocolate cake is certainly *more* tasty to them than apples, apples

occurrently judging that *p*, and therefore that M (when activated) realizes occurrently judging that p. But, in order to rebuff examples of people who appeared to have M activated but also appeared to not occurrently judge that p, the author claimed that M was in those instances misfiring or failing to play its role. I submit you would want to know more about M to know how to assess such a rebuff: You would want to know what it normally looks like, at the neurological level, when activated; how it normally interacts with other neurological mechanisms of relevantly similar kinds; and, of course, what it looks like in the cases at hand. You would not be satisfied with some symptoms of this mechanism being activated, such as that the person typically has an elevated heart rate, slightly dilated pupils, and can be seen scratching her head.

are still tasty enough. Deciding in this way is faster and cheaper than full-blown deliberation, which would involve surveying all the items in the cafeteria and weighing them against one another.[9]

To spell this idea out further, I suggest that we understand these shallow cognitive processes as realizing *dispositions to decide by using a heuristic* (hereafter, heuristic dispositions or simply HDs). In the above case, the suggestion might be that the employees are disposed to decide by using the Satisficing Heuristic: Choose the first option that is presented that exhibits the relevant feature to some satisfactory degree (Simon 1956).

To give another example, consider the well-known "default effect": Evidence indicates that people are more likely to choose an option if it is somehow presented as the default choice, perhaps because the digital interface they are using has that option selected, but allows them to change their selection to a different option.[10] We might explain this by hypothesizing that people are generally disposed to decide by following the Status Quo Heuristic: If you are not able to make an informed decision given the information you have, do nothing.[11]

There are many other such heuristics and, accordingly, dispositions to decide by using them. Which such dispositions *we* possess is a topic of investigation for behavioral psychologists who are interested in understanding our agency in these terms (e.g., Gigerenzer & Selten 2001; Payne, Bettman, & Johnson 1993). Let me give a few more examples of well-known heuristics from Gigerenzer and Gaissmaier (2011), adapted for the domain of practical choice. There is a variety of HDs that people may have and so a variety of ways through which nudging might work on us.

One example is the so-called Take the Best Heuristic. To use this heuristic, an agent must first have a well-defined choice set, and must be able to rank the importance of various features that the options in that choice set might have. For example, if the choice is between which car to buy, the features might be safety rating, price, comfort, top speed, etc., and these are features that the agent can

9. Contrary to Thaler and Sunstein, Saghai's characterization, which I follow, allows that these processes might be carried out consciously, and they need not be error-prone in all or most situations (cf. Thaler & Sunstein 2008: 17–39; see also Gigerenzer 2015). This issue will return in Section 3.1. Finally, note that Levy partly defends nudging on the grounds that the distinction between what I consider to be heuristic reasoning, and full-blown practical deliberation, is specious (2019: §3.2). While I am sympathetic to this claim, the arguments of the present paper do not depend on it.

10. For some evidence that such defaults are effective, see van Dalen and Henkins (2014), Johnson and Goldstein (2003), and Li, Hawley, and Schnier (2013). For a recent meta-analysis, see Jachimowicz, Duncan, Weber, and Johnson (2019).

11. Samuelson and Zeckhauser (1988) say that behavior that accords with such a heuristic exhibits a *status quo bias*. That may be so, but calling it a 'bias' suggests taking a stand on some of the issues I will tackle in the remainder of this paper; at present, I will to be as neutral as possible on the *quality* of thinking and deciding that follows such a heuristic.

rank in terms of relative importance. Additionally, these features are viewed as having only two values: For any feature *f*, either an option has *f* or does not—or, failing that, the agent can use thresholds to generate these binary values. Then, the agent selects the most important feature that discriminates among the options (i.e., some of them have this feature and some do not). If the remaining options with that feature are precisely one, she chooses that option; if it is not, she restarts the process on that smaller subset of options using the next-most important feature that discriminates, and so on, until one option remains.

Similar to the Take the Best Heuristic is the Tallying Heuristic. Here, again, all features that are relevant to the choice are viewed as having discrete values, though they can range from negative to positive—most naturally, the set of values is [-1, 1]. The agent then sums up these values for each option and chooses the option with the greatest positive sum.

These are just a few of the many heuristics that behavioral scientists hypothesize that we are disposed to use in making decisions. The general contours of heuristics should be clear, however: They are rules that lead to specific decisions on the basis of a circumscribed set of considerations. And understanding nudging as operating via *dispositions* to decide using such heuristics (HDs) allows us to analyze the phenomenon further than Saghai's account and to link nudging explicitly with heuristics such as these. This analysis will also get several interrelated ethical worries about nudging into clear view, which I will turn to in Sections 2 and 3.

1.3. Enabling, Triggering, and Manifesting: The Components of Nudging

So let me now develop the analysis further. I will focus on the Status Quo Heuristic due to its simplicity—following it is a one-step process—to suss out the general features of HDs. The first thing to notice about the Status Quo Heuristic is that there are only some situations in which essential prerequisites for using the heuristic are met: An agent must be able to form an opinion about whether her information is sufficient to be the basis of an informed choice, and she must also be able to do nothing (viz., she must not be in a forced-choice situation). Accordingly, a disposition to decide by using this heuristic has certain *enabling conditions*, namely, conditions under which the essential prerequisites of using the heuristic are met.

The second thing to notice about the Status Quo Heuristic is that there are only some situations, even among those in which its essential prerequisites are met, in which it *will be* followed, all else equal. Namely, it is only in those situations in which the agent (who is disposed to decide by using the heuristic)

concludes that she lacks sufficient information to make an informed choice, and all else is equal, that she uses the heuristic and does nothing (viz., makes no choice). Other agents might conclude that they have sufficient information, and then proceed to try to make an informed choice. What these possibilities reveal is that a disposition to decide with the Status Quo Heuristic has certain *triggering conditions*, namely, conditions under which the heuristic in question is actually used, all else equal.

The above 'all else equal' qualifier allows us to distinguish a third important feature of HDs. As I just said, an agent might conclude she lacks sufficient information to make an informed choice, and then use the Status Quo Heuristic; while another might conclude that she has such information and try to make an informed choice some other way. But, of course, an agent might conclude that she lacks sufficient information, but for other reasons proceed to try to make a choice in some other way—thus not using the Status Quo Heuristic, which tells her to do nothing in that case. For all else might not be equal by her lights. Cases like these show that the triggering of an HD does not entail the *manifestation* of that HD in the form of a choice.

These internal features of HDs allow us to distinguish a feature of an agent's choice situation *being a nudge,* from the agent herself *being nudged*, and from an agent performing a *nudged action*. In particular:

> *Definition of a nudge*. A feature *f* of a choice situation *s* is *a nudge* for an agent *A* just in case *f* enables an HD that *A* has.
>
> *Definition of being nudged*. An agent *A is nudged* by a feature *f* of choice situation *s* just in case *f* is a nudge for *A* and her HD, which *f* enables, is triggered in *s*.
>
> *Definition of nudged action*. An agent *A* performs a *nudged action φ* in choice situation *s* just in case she is nudged by a feature *f* of *s* and her triggered HD, which *f* enables, manifests in a decision that issues in her *φ*ing.

1.4. Clarifications and Discussion of Putative Nudge Strategies

These definitions require a few remarks. First, on the definition I've given of a nudge, not *everything* is a nudge for agents whenever it affects them in some way: Many features of the environment are features that the deciding agent merely takes into account, either when using a specific heuristic or engaging in more full-blown deliberation. To see this, look at taxes. A tax on a good doesn't plausibly enable a specific HD in ordinary agents, because the HDs we plausibly have

from both an evolutionary and a psychological point of view are not catered to making decisions in tax-laden environments. Nevertheless, in using specific heuristics or engaging in full-blown deliberation, the reasons for and against any particular choice will include those created by the presence or absence of the taxes that exist in that situation.[12]

Second, and with that being said, many things will count as nudges and, accordingly, nudges can have many sources—natural limitations, information-processing capacities, other agents, and so on. In particular, my definition of a nudge allows that nudges might be present in a choice situation without being there *by design*—that is, without having been intentionally put there by a designer. Similarly, my definition of being nudged allows that an agent might be nudged without anyone else intentionally nudging her. To be sure, such interpersonal dynamics are surely relevant to the ethics of nudging: We want to know whether nudging, as a species of interpersonal influence, is objectionable. As we have seen already, most theorists enfold this interpersonal aspect into the very definition of nudging.

Now, my analysis should be easy to extend in this direction: We can define what it is for one agent A to nudge another agent B in terms of A's implementing a nudge and/or bringing it about that B is nudged. In my view, however, extending my analysis in this way is secondary to my aims. As we will see in Section 2, the ethics of nudging one another hangs on the rational status of the processes through which nudging works: In nudging one another, it matters whether we are tapping into processes that are not rational. So the rational status of these processes is prior to the ethics of influencing one another through them. And this rational status does not itself hang on whether nudging is by design: It hangs on how those processes themselves work, along with the nature of rationality more generally. In this respect, to focus on nudging by design is to put the cart before the horse.[13]

Third, when an agent is nudged, it is not necessary for *the nudge* to trigger an HD the agent has. This is as it should be. Consider a default nudge again, along with the disposition to decide by using the Status Quo Heuristic, which is plausibly the disposition that the nudge enables. When it is triggered, this disposition is triggered by the agent's assessment of her information as inadequate

12. Interestingly enough, one could imagine a case in which a tax is a nudge. Imagine that a particular agent has a visceral hatred of taxation (believing it to be, in every case, unjustified theft), and that this hatred has cultivated in her a kind of automatic, negative response to taxes—perhaps she refuses to buy any goods that have consumption taxes attached. For her, the consumption tax might act as a nudge, because she plausibly is disposed to decide by using the following heuristic: *If a good has a consumption tax attached, do not choose it*. For her, the tax doesn't stand as a disincentive, or cost, associated with the good; the tax excludes the good from consideration entirely. My definition of a nudge gives us the conceptual tools to make and explain exactly this distinction.

13. My thanks to Barry Maguire and an anonymous reviewer for pressing me to address this.

to make an informed decision, not the fact that she is given a default option (i.e., the nudge itself). To be sure, the nudge itself remains explanatorily relevant, but not by being the proximate cause of the intention to do nothing in this choice situation.

Fourth and finally, on these definitions of nudges and being nudged, it is possible that an HD that an agent has will be triggered (and she will thus be nudged), without it following that she decides in favor of the option she is being nudged to choose, or that she executes that decision in action. This will matter in Section 3.

This analysis helps us make some principled discriminations among strategies of influence that sometimes traffic under the heading of "nudging".[14] I have discussed the default effect already, and we can see that whether pre-set defaults count as nudges hangs on whether agents who encounter them are disposed to decide using the Status Quo Heuristic (or at least some relevantly similar heuristic), and on whether the default enables some such disposition.

Many of Thaler and Sunstein's original examples of "nudging" have been challenged as not really being nudging at all. For example, Thaler and Sunstein claim that educational campaigns about rates of tax compliance and rates of binge drinking count as nudges to, respectively, pay one's due taxes or binge drink less (2008: 66–68). They also claim that including a smiley face on a household's energy bill that corresponds to how one's household energy consumption compares to the local average—green and smiling for less-than-average consumption, red and frowning for more-than-average consumption—counts as a nudge to consume less energy (2008: 68–69). However, it is not at all clear why an educational campaign should count as a nudge; after all, providing information to someone in the hopes that they will deem it relevant to their choice, in the way you hope, appears to be the same kind of influence as persuasion (Hausman & Welch 2010: 127).[15] My analysis can now explain why this should not count as a nudge: The mere provision of information does not, in itself, enable any disposition to decide using a specific heuristic.

The latter strategy—smiley faces on energy bills—presents a more subtle case. On the one hand, the smiley faces convey information about the approval or disapproval one's energy consumption does (or would) receive, at least from environmentally conscious bureaucrats. That is some reason to think this design feature works at least partly through the provision of information. On the other

14. My thanks to an anonymous reviewer for pressing me to discuss specific (putative) nudging strategies.

15. That is not to say that these campaigns might not, for all that, still be manipulative. We can manipulate people through the tactics of persuasion, after all. The question at hand is whether these particular tactics should be categorized as the same kind of influence as persuasion, or as nudging.

hand, however, it is plausible that the smiley faces induce affective responses in the people who receive them, which shapes their attention going forward and motivates them to consume less energy. So chalking this strategy up to straightforward provision of information might be too simple.[16]

Now, our intuitions on these cases ought not be treated as entirely dispositive, since, as I pointed out in the introduction, "nudging" and its cognates are novel terms of art in the present context. As I outlined in Section 1.1, the term is meant to capture a distinctive kind of influence, which previous researchers have done largely by appealing to what nudging is *not*: neither force, nor incentivization, nor persuasion. However, as I also argued there, this differentiation is not enough to assess the rational status of the processes underlying nudging.[17]

We can see this quite forcefully with the present example. Suppose it is true that the smiling and frowning faces induce attention-directing and motivating affective responses in the people who receive them on their energy bill, and that these underlying psychological processes render the influence neither force, nor incentivization, nor persuasion. Now, there are many other cases of influence that work via diverse psychological processes, such that the influence is neither force, nor incentivization, nor persuasion. For example, just as we might shape someone's attention and motivation by inducing certain affective responses, we might stoke their imagination, gain their trust, or massage their mood. The problem is that we do not have any reason to suppose that these diverse, underlying psychological processes will be on a par with respect to their rational status in general.

For this reason, it will not do to treat all such cases alike—that is, to treat any influence that fails to amount to force, incentivization, or persuasion as a case of nudging, and then proceed to ask after the rational status of the processes implicated in nudging. So what is needed is a more discriminating account of nudging, such as the one I offer. With respect to hard cases such as smiley faces on energy bills, my analysis tells us what it would take for such features to count as nudges: The inclusion of these smiley faces would have to enable a disposition to decide using a specific heuristic.[18]

More generally, this is the kind of bar some purported nudge must meet to count as a nudge. To be a nudge (relative to some agent), a feature of a choice situation must enable a disposition (in that agent) to decide by using a specific heuristic, which is a matter of that feature's meeting some essential prerequisites

16. For a similar point, see Selinger and Whyte's discussion of "fuzzy nudges" (2011: 927–28).

17. Where the assessment of their rational status matters ethically. See Section 2.

18. My thanks to an anonymous reviewer for presenting another hard case similar to the one under discussion, which prompted me to expand my response in fruitful ways. That person presented to me a case of attaching icons of the human eye at worksites where disregarding safety protocols was likely to lead to injuries.

for that disposition to be triggered or manifested.[19] Typically, the mere provision, or making salient, of specific considerations will not meet this bar. This is because, as we have seen, heuristics for deciding lead to specific decisions on the basis of a circumscribed set of considerations; the relevant considerations are not essential prerequisites for using those heuristics, but rather figure into the use itself.

2. Reasoning and Self-Guidance: Getting a Worry about Nudging into Focus

Recall that there is another crucial component to Saghai's conception of a nudge: It must be "substantially noncontrolling", that is, "B can easily resist A's attempt to get her to x" (2013b: 499, variables changed). What this condition of substantial noncontrol amounts to is a difficult question in need of further theoretical work. After all, many worry that nudging undermines an agent's ability to guide her own actions, which is a key component of *autonomy* and therefore something that ought not be undermined without good reason. For such thinkers, the benefits of nudging and the value of protecting, preserving, or promoting autonomy present a trade-off that must be handled with care. For example, Luc Bovens writes

> There is something less than fully autonomous about the patterns of decision-making that [nudging] taps into. When we are subject to the mechanisms that are studied in 'the science of choice', then we are not fully in control of our actions. (2009: 209)

Similarly, Hausman and Welch claim that, "[W]hen [nudging] does not take the form of rational persuasion, [nudged agents'] autonomy—the extent to which they have control over their own evaluations and deliberation—is diminished" (2010: 128).

I suggest we interpret the worry here in the following way. Because nudging is a matter of enabling, triggering, and manifesting dispositions to decide by using heuristics, as opposed to choosing via full-blown deliberation, nudging undermines an agent's ability to guide her own actions by affecting her reason-

19. Typically, the dispositions agents possess are treated as a feature of their psychological endowment, prior to their being in the choice situation in question. However, an interesting and underexplored question is whether designers can nudge us in two steps, first by *providing* the relevant heuristic, and then designing the situation such that the situation enables and triggers a disposition to decide with that heuristic. In my view, designers evidently *can* do this because they *do* do this. Among other things, this two-step process is part of what it is to design a *gamified* artifact.

ing in certain detrimental ways. I will spell this out more in a moment, but let me emphasize how it gives rise to a distinctly ethical objection to nudging. If nudging undermines an agent's self-guidance, then, when an agent A influences another agent B via nudging, A exerts substantially *controlling* influence on B, and this controlling influence is pro tanto wrong.

Thus, it turns out to matter whether nudging really undermines a nudged agent's self-guidance; and so, in turn, whether one agent A's influencing another agent B via nudging really can be substantially noncontrolling in Saghai's sense.

To be clear, this worry about self-guidance need not be founded on the suspicion that nudged agents *do not reason at all* while being nudged. As the reader can no doubt guess from what I say in Section 1, I see no principled reason to suspect that nudging forecloses entirely on reasoning. In fact, the analysis I offer there suggests that nudging works *through* reasoning with heuristics, namely, by utilizing dispositions the agent has for deciding on the basis of limited, cognitively cheap assessments of her options. Deciding on the basis of assessment of one's options is, I take it, practical reasoning stripped down to its skeleton.

Rather, I suggest that the worry relies on the idea that agents do not reason *well* while being nudged, since reasoning well is typically associated with careful and exhaustive deliberation, not making limited, cognitively cheap assessments of one's options. Furthermore, the worry is that this in turn undermines an agent's ability to guide herself in action.

Here's a plausible picture to motivate this connection. First, an agent's actions are self-guided only to the extent that she is a unified agent.[20] This is because self-guided actions are actions guided by the agent in light of her practical point of view on the matter at hand; were the agent in some serious ways disunified, she would not have a determinate practical point of view at all, but rather a variety of different takes on the matter at hand, none of them having any claim to being distinctly *hers*. Second, an agent is unified only to the extent that she reasons well. Reasoning well involves, among other things, regulating our application of various domain-specific rules in light of others, with the more general aim of bringing our beliefs, intentions, and so forth into global consistency and coherence.[21] Reasoning well is how we manage our mental economy in general, which is essential to being a unified agent because it is the process whereby we put the mess inside into some kind of intelligible order of epistemic and voli-

20. Here and in what follows, the locution "only to the extent that" is meant to track a necessity, not sufficiency, relation, though one that obtains between statuses that one can exhibit to varying degrees. It is meant to allow, for example, cases in which an agent is *not* self-guided to the same extent that she is unified. But it disallows cases in which her actions *are* self-guided to a *greater* extent than she is unified.

21. Failing to regulate ourselves in this way may still count as reasoning, but will be reasoning in a way that follows rules somewhat blindly, without sensitivity to what one believes, intends, and so forth beyond what the rule specifically operates on.

tional commitments. And so it is, in turn, a key feature of self-guidance, too. So, putting these pieces together, if nudging an agent prevents her from reasoning well while she is nudged, this will undermine her unity and hence self-guidance.

Given the analysis of nudging offered in Section 1, and this sketch of self-guidance, we can state more clearly what the worry is. Recall that nudging has three parts: the nudge itself, the agent being nudged, and the agent performing a nudged action. These are a matter of, respectively, enabling, triggering, and manifesting an HD that the agent has. So the worry, I suggest, is that nudged actions are not self-guided actions when (and because) being nudged prevents the agent from reasoning well about what she shall do.[22]

With this clarification on the table, we are now in a position to understand one way in which the *rational status* of the underlying psychological processes matters for the ethics of nudging, as I mentioned at the outset of this paper. Section 1 affords a dispositional analysis of the psychological processes through which nudging works: They realize dispositions to decide using heuristics. So a slightly different slant on the worry I have just given goes like this. If these psychological processes realize procedures that are not rational, then, when a choice situation has a nudge, and an agent in that situation is nudged accordingly, she is being influenced to decide in a non-rational way. Deciding in a non-rational way undermines the agent's guidance of her own actions.[23]

This worry is suitably general to capture a lot of the unease around nudges that we have seen in this paper. In the next section, I'll show what must be true of particular instances of being nudged for this worry to be defused. Since this

22. There's a very close-by worry that I will not address in the main text. It is possible that, *even if* an agent can reason well about what she shall do while being nudged, she'll do the nudged action whether or not her reasoning well concludes with the intention to do that action; and this might give rise to the worry that nudging undermines self-guidance by rendering the agent's good reasoning causally otiose. As I said, I won't address this worry in much depth here, but let me just say one thing to defuse the worry a bit. As I mentioned before, I think there is no reason to think that nudging bypasses reasoning entirely, and I showed in Section 1 that we get a comprehensive and elegant analysis of nudging if we think about nudging as working through heuristic reasoning. If nudging works through reasoning in this way, then reasoning per se cannot be causally otiose when nudging is happening. But if *that's* right, it is not clear what reason we have for thinking that reasoning *well* is causally otiose when nudging is happening. This is because the difference between reasoning per se and reasoning well is a normative one, where token instances of either are realized by the very same type of non-normative, sub-personal processes.

23. There are, to be sure, other ways in which the rational status of these procedures might matter for the ethics of nudging. For example, if these procedures are not rational, then when we try to influence one another by way of them, we might fail to *treat* one another as rational; and this would explain one way in which intentionally nudging one another is wrong. For more on this, see Schmidt's 'Weaker Argument' against this idea (2019: 520–27), and Rozeboom for commentary on how the rational status of these procedures, and treating one another as rational, can come apart (2020: especially 110–11). My thanks to an anonymous reviewer for encouraging me to be more explicit about how rationality and self-guidance relate with respect to the worry I focus on.

defense of nudging need not apply to all nudging whatsoever, the reader should note that the ambitions of this part of the paper are relatively modest. In the section after that, I'll discuss how defusing this worry vis-à-vis particular instances of being nudged has an interesting consequence: It crystallizes a distinct, more perplexing worry about nudges having to do with manipulation.

3. Reasoning Well while Being Nudged

To begin, I need to say more about reasoning well. But I will not take up the arduous task of giving reasoning well (or just reasoning, for that matter) a complete analysis, nor even that of saying what makes an instance of reasoning good.[24] I will only highlight some of its characteristic marks. Consider that good reasoning can be characterized as using rules such as the following:

> *Means-End Coherence Rule*. If you intend to φ, and believe that you must intend to ψ in order to φ, then intend to ψ.
>
> *Expert Testimony Rule*. If you believe that a relevant expert *e* said that *p*, then believe that *p*.

Good reasoning characteristically proceeds by following rules such as these (as well as many others). This much should be uncontroversial for present purposes. Even if a person suspected that nudges enabled following rules that were *not* characteristic of good reasoning, that is little reason to suspect that *these* rules are not characteristic of good reasoning.

However, reasoning well must still exhibit a considerable degree of *flexibility* vis-à-vis these good rules. Consider the Means-End Coherence Rule. A reasoner following this rule might reason from her intention to make peanut brittle this weekend, and from her belief that she will have to intend to (and to actually) buy peanuts at the grocery store before then, to the intention to buy peanuts. However, intending to buy peanuts might strike her, for independent reasons, as a really bad idea—perhaps because her mother, who is visiting this week, has an extremely bad peanut allergy. And so, reasoning well, she will reconsider and drop her intention to make peanut brittle this weekend and never take up the intention to buy peanuts. Now consider the Expert Testimony Rule. A reasoner following this rule might reason from her belief that she heard Dr. Smog, an expert on the health effects of tobacco-use, say on national TV that smoking

24. For some work on those questions, see Boghossian (2014), Broome (2013), McHugh and Way (2018a; 2018b), and Wedgwood (2006; 2012).

doesn't cause lung cancer, to the belief that smoking doesn't cause lung cancer. However, she might for independent reasons think that Dr. Smog's salary is being paid by Philip Morris USA (the manufacturer of Marlboro cigarettes), and so refrain from coming to believe that smoking doesn't cause lung cancer.

The particular psychological dynamics of these two cases are interestingly different, but I do not wish to linger on those differences here. The point I am illustrating is simply that, even when it comes to rules that it is generally good to follow in reasoning well, the reasoner herself still has to be rather flexible with respect to those rules if she is to reason well.

These characteristics of reasoning well provide us with two bars that an agent must clear if she is to count as reasoning well when being nudged. First, the rules she follows in being nudged have to be *good* rules to reason with. Second, while being nudged, she has to be able to be *flexible*, in the ways illustrated above, with respect to those good rules.

3.1. Are the Heuristics Involved Good Rules to Reason With?

It might seem that the heuristics that nudges enable us to use, such as the Satisficing Heuristic, Status Quo Heuristic, Take the Best Heuristic, or the Tallying Heuristic, are *prima facie* not good rules to follow when reasoning. This seems to be suggested by their fast and frugal nature: Being cognitively cheap to follow, and involving a circumscribed set of considerations, it seems likely that they will be systematically biased in the responses they license. The Satisficing Heuristic, once again, serves as a nice illustration: Because following it involves a sequential search through options that terminates once a satisfactory option has been found, choices made by following it will be biased toward satisfactory options discovered earlier rather than whichever option would be considered best by the agent's own lights, were she to engage in some more thorough procedure of evaluation that takes into account all relevant considerations.

The most prominent proponents of nudging, Richard Thaler and Cass Sunstein, themselves endorse and repeat this skepticism about the quality of the heuristics that nudges enable. For example, they say that in thinking in ways that use these heuristics, people "make pretty bad decisions" (2008: 5), are "clueless" (2008: 19), listen to "the lizard inside" (2008: 22), and are more like Homer Simpson than Mr. Spock (2008: 22)—and that is all before chapter 2 of their agenda-setting book. That people are allegedly befuddled so easily and across so many areas of life plays a key role in their justification for nudging in the first place, because nudges could (in theory) be designed to redirect these biased ways of choosing toward the outcomes that would be deemed best by the agent herself, were she not so befuddled (2008: 5). Given the worry I am investigating,

it should be no surprise that concerns about the autonomy-undermining power of nudging have been so durable. On this front, at least, they have not been the best advocates for their own agenda.[25]

In the background of this skepticism is a certain conception of *procedural rationality* in light of which deciding by using heuristics is not procedurally rational and, as a result, doing so will not amount to reasoning well.[26] Generalizing a bit, the worry might be that using heuristics to decide is not procedurally rational for two reasons. First, heuristics lead to decisions on the basis of circumscribed sets of considerations, rather than on the basis of any and all considerations that are available to the agent and relevant to the decision. Second, the heuristic cannot lead to decisions that are responsive to the genuine normative force of all of those available and relevant considerations (since it is not even based on all of them).

To fully address this worry would require a lengthy treatment of procedural rationality, which I cannot provide here. But let me push back just a bit. I take it that whether a rule of reasoning is a *good* rule for a situation depends on the limitations and constraints of the agent who might use it in that situation. More to the point, a rule of reasoning is *good* relative to an agent and a situation only if she *can* use it to reason in that situation. The conception of procedural rationality

---

25. An anonymous reviewer has claimed that this characterization is disingenuous. According to that reviewer, Thaler and Sunstein's criterion for whether an agent's decision is bad is whether or not it makes the agent better off, not anything about the rules with which she decided. But this is false. Here is the relevant passage in full:

> Drawing on some well-established findings in social science, we show that in many cases, individuals make pretty bad decisions—decisions they would not have made if they had paid full attention and possessed complete information, unlimited cognitive abilities, and complete self-control. (Thaler & Sunstein 2008: 5)

It is true that this passage occurs within a broader context in which Thaler and Sunstein are outlining their "libertarian paternalist" agenda, which involves nudging people to make them better off. It is important, however, not to conflate Thaler and Sunstein's negative assessment of how people typically reason with their libertarian-paternalist justification for nudging *in light of* that assessment. Their general assessment of human reasoning is bleak, and it expresses the skepticism I am highlighting in the main text (see Thaler and Sunstein: ch. 1 for the longer version). Their justification for nudging is posterior to this assessment: *Given* the poor quality of people's reasoning, intervening by nudging is justified when and because it makes those people better off than they would be were they left to their own devices.

26. This normative conception is typically associated with, though by no means unique to, the "heuristics and biases" research program in psychology and economics, which descends from the experiments of Daniel Kahneman and Amos Tversky. See Tversky and Kahneman (1973; 1974) for two classic papers in this tradition, along with Kahneman, Slovic, and Tversky (1982) for a compendium of studies in this vein; and see Kahneman (2011) for a synoptic summary of this program that reflects its development and eventual integration of "Dual Systems" theory of mind. It is also, evidently, the conception lying behind Thaler and Sunstein's own views, which are explicitly informed by the heuristics and biases program. My thanks to an anonymous reviewer for pressing me to contextualize this issue.

that we have before us is hard to square with this. It is not clear that, in general, we can reason with rules that are based on, and responsive to the normative force of, all available and relevant considerations in the various situations in which we find ourselves. To bring to mind all of the available and relevant considerations, and to respond to the normative force of those considerations, would require considerable working memory, adequate time, and considerable computational resources. I am skeptical that, in general, we possess these capacities to the requisite degree. But, in any case, we often enough cannot bring them to bear in the situation at hand. Procedural rationality must give its stamp of approval to rules that we can actually use in the course of living our lives.

Careful attention to the quality of particular heuristics, as used by reasoners like us across a variety of situations, is a more grounded approach.[27] Here are a couple of illustrations from which we can draw more general lessons. When using the Tallying Rule with only weights of -1 or +1 associated with each relevant feature, predictions about the overall quality of an option become better than multiple regression analysis as, among other things, the number of relevant features increases, and the average correlation between those features increases (Goldstein et al. 2001: 175–76, themselves drawing on Einhorn & Hogarth 1975). And following the Status Quo Rule is better than trying to decide on the basis of one's own information about the options when the default option conveys an implicit recommendation by the expert who put it in place.[28]

I take these sorts of illustrations to be compelling: Following heuristic rules like these can produce quite good outcomes in situations like those marked by considerable complexity, uncertainty, and reliance on others. This gives us some reason to think that heuristics can be good rules for reasoning, *provided that* the situation has the right sort of features.[29] Notice, importantly, that a nudge need not foreclose on this provision being met: It need not make the situation one in which the heuristic is not a good rule to follow. Indeed, a well-meaning nudge designer can implement nudges that enable dispositions to reason with heuristics that *are* good rules to reason with in that situation, considered independently of the nudge itself.[30]

27. Predictably, this approach is associated with conceptions of rationality that rival the one just discussed in the main text—most commonly, *ecological* conceptions. Gerd Gigerenzer and colleagues are most known for advancing this conception with their "fast and frugal heuristics" program. For a good overview, see Gigerenzer and Selten (2001), and see Simon (1957) for an important and influential precursor. For other philosophical deployments of this conception, see Morton (2011; 2017) and Schmidt (2019).

28. This is an expansion on Gigerenzer's discussion of framing effects, of which defaults are one species (Houk 2019: 415; Levy 2019: 290; see Gigerenzer 2015: 367–69).

29. My thanks to an anonymous reviewer for encouraging me to be explicit about this provision.

30. To be sure, there are additional ethical issues here that depend on the nudge designer's alternatives, especially if she has the option of not intentionally implementing any nudge at all.

There is, however, another source of pessimism that merits response. One might think that a heuristic can't be a good rule to follow if it bears no relation to what the agent herself takes her reasons to be. The heuristic itself would strike the agent, if she were to consider it, as arbitrary. This is easiest to illustrate with the "ballot order effect": Candidates higher up on an electoral ballot (especially first) often receive a statistically significant bump in vote share (Alvarez, Hasen, & Sinclair 2006; Ho & Imai 2006; 2008; Koppell & Steen 2004; Krosnick, Miller, & Tichy 2004; Lutz 2010; Marcinkiewicz 2014; Meredith & Salant 2013; Miller & Krosnick 1998; Pasek, Schneider, Krosnick, Tahk, Ophir, & Milligan 2014; and Webber, Rallings, Borisyuk, & Thrasher 2014). Now, suppose that this effect was best explained by the fact that many such voters chose the higher-listed candidates by following the simple heuristic: Choose the option (candidate) listed first.[31] However, if they were to actually consider whether the fact that a candidate was listed first was a *reason* to choose her, they would likely deny that it is; accordingly, the corresponding heuristic would likely strike them as arbitrary. Finally, suppose that, for some reason unbeknownst to these voters themselves, the ballot order of candidates really *is* correlated with their quality. That would not change the fact that this simple heuristic, of preferring the first candidate, is not a good rule to reason with in this situation.

To put the point more concisely: Even if using the relevant heuristic produces quite good choices in the situation at hand, it might not be a good rule *for the agent* to reason with in that situation when and because she does not see herself as having reasons to respond in the way that heuristic licenses.[32]

My response to this is, in the main, conciliatory. However, let me make a few important qualifications. First, cases like this one do not plausibly generalize across all instances of being nudged, since many of the relevant heuris-

---

Here, whether the situation will nonetheless have nudges anyway is a relevant consideration, as is whether the would-be nudged agent could effectively engage in full-blown deliberation in that situation. For discussion on these matters, consult, for example, Engelen (2019: 220–23), Houk (2019: 416–17), and Sunstein (2015: 420–22). The issue I'm considering at present is independent of these questions. These questions are ethically relevant whether or not it is pro tanto wrong for one agent to control another via nudging.

31. It is worth noting that the best explanation of this effect is not necessarily just that the voters choose by following such a crude rule. For example, they may proceed down the list in order, searching for reasons they can recall to choose each candidate, but this search often terminates before the entire list is considered due to fatigue and strain on working memory. And how long this search goes on might be a function of the individual voter's fatigue, working memory capacities, and so forth. But, for all the variance among those features, we would expect, at the population level, to see the progressively further-down candidates considered by progressively fewer voters trying to decide in this way. See Krosnick et al. (2004) for a discussion of this and other possible explanations.

32. Cf. Doris (2015; for a brief overview, see Doris 2018). However, for Doris, the fact that the person chooses on the basis of a consideration she does not herself take to be a reason threatens to defeat her *morally responsible agency*, not the quality of her reasoning *per se*.

tics directly involve the agent's own normative point of view. To return to our main two cases, it appears that following the Status Quo Heuristic requires an assessment about the sufficiency of one's own information vis-à-vis making the choice; and this is plausibly a matter of assessing whether one's body of information provides one with sufficient reason to choose any option in particular. And following the Satisficing Heuristic requires assessing whether the option one has just encountered is satisfactory with respect to what matters; and this is plausibly a matter of assessing whether there is sufficient reason to choose it. The other rules I have mentioned, too, like the Take the Best Heuristic and the Tallying Heuristic, require evaluating the quality of one's options along several dimensions, and therefore plausibly involve one's take on the reasons for and against them.[33]

Moreover, it is not entirely clear that heuristics that do not involve the agent's normative point of view directly cannot still be good rules to reason with, provided that the agent is able to exhibit flexibility with respect to them (see Section 3.2).[34] Still, the point here is a good one that can be incorporated into a 'best practices' for designing nudges: When it comes to respecting the self-guidance of agents, nudges should enable heuristic dispositions (HDs) that deploy the agent's own normative point of view.[35]

---

33. Note that there are important empirical questions here. To return to our earlier example, does the Wellness Guru's cafeteria nudge work on us (to the extent that it does) by enabling a disposition to follow the Satisficing Heuristic, or a simpler heuristic, like *take the first option presented*? The former plausibly involves one's own normative point of view, the latter not so much. For some commentary on how cognitive scientists attempt to answer such questions, see Hutchinson and Gigerenzer (2005: Section 4).

34. With order effects on electoral ballots, there's some suggestive evidence that they can, for ballot order effects are larger in cases where voters plausibly have little information or are ambivalent about their choice—specifically in elections for offices of low visibility, or in which turnout was higher, or in which the race was not very close. See, e.g., Alvarez, Hasen, and Sinclair (2006), Ho and Imai (2006; 2008), Koppell and Steen (2004), Miller and Krosnick (1998), and Pasek et al. (2014).

35. There are interesting questions here regarding the ethics of education. It might be that teaching someone well sometimes requires getting them to follow rules that do involve their actual normative point of view, perhaps in part with the goal of developing and improving that point of view itself. For example, in teaching students critical thinking, it is often necessary to get them to stop affirming the consequent when reasoning deductively. The problem is precisely that their normative point of view is getting things wrong: The truth of the consequent of a material conditional is not a reason to believe the antecedent of that conditional, though they erroneously take it to be such a reason. So, in teaching them to think better, it is plausible that we have to get them to follow rules that run antithetical to their own normative point of view. In formalized educational contexts, one might think this is straightforwardly compatible with respecting their self-guidance when and because they consent to being directed by their teacher in this way, but this is not entirely clear. After all, it is not clear that consent can be given in the requisite sense because, *ex hypothesi*, the student cannot know in advance exactly *what* she is consenting to—what directions she will be given and expected to follow—because she does not know in advance the quality of the rules of reasoning she already follows that will be subject to such corrective

3.2. Can the Agent Be Flexible with Respect to Those Heuristics?

The case made in the previous section tees up a second, natural skepticism about the quality of reasoning with heuristics. The case there was that using these heuristics produces very good choices under various common conditions, such as uncertainty or when recommendations are being pragmatically conveyed. Still, these conditions are not always met, and, even when they are met, that doesn't mean that the HD that happens to get triggered in the nudged agent is a good one for that situation. This is again easiest to illustrate with the disposition to decide with the Status Quo Heuristic. As I suggested a moment ago, this can be a good rule to reason with when an option has been made the default to convey an implicit expert recommendation to the agent about what to choose. But not all defaults convey implicit recommendations by experts—they might rather be recommendations given by other, self-interested parties, or simply a quirk in the presentation of the options. In *these* situations, deciding with the Status Quo Heuristic will plausibly not be reasoning well.

Zooming out a bit, this skepticism is a species of a larger skepticism about dispositions to decide with heuristics. Namely, the skepticism is that these HDs are too *inflexible* to realize good reasoning since they can produce only a very limited range of choices in response to situations that must be structured in rather specific ways, whereas good reasoning is typically thought to be characterized by its immense *flexibility*, both in its potential outcomes and the situations that can prompt it.

To begin to address this, let me elaborate on a point I made at the end of Section 1. Dispositions are defeasible in that their being triggered does not necessitate their being manifested. (This is typically captured by an 'all else equal' qualifier when describing them using conditionals.) So, since being nudged is a matter of an agent's HD being triggered, it is possible that this disposition will be interrupted and so its manifestation not brought about. This means that there is nothing about being nudged *per se* that gives us reason to think that the nudged action—the manifestation of the relevant HD in action—*must* be performed upon being nudged.

Now, while being nudged can in this thin sense be interrupted, that is not to say that the agent herself can be flexible in her reasoning with respect to the heuristics she is disposed to use. However, this does point to what it would be for her to be flexible in this way: At least some relevant class of defeaters for these HDs would have to be attributable to *her*.

---

direction. The issue seems to me even less clear in other contexts. These issues bear, ultimately, on the manipulation worry I spell out in Section 4, but it is beyond the scope of this paper to give them adequate treatment.

Here again we can return to the idea broached in Section 3.1. The class of defeaters that I propose are attributable to the agent herself are those considerations that she takes to be reasons to do something other than use the relevant heuristic.[36] It seems to me that what an agent takes to be a reason can be a defeater in this sense in at least two kinds of case.

The first kind of case are those in which the agent takes herself to have *greater* reason to choose differently than the HD disposes her to, however much reason she takes herself to have to choose in accord with the heuristic involved in that HD. The second are those in which the agent takes herself to have *exclusionary* reason to choose differently. Exclusionary reasons to do otherwise function slightly differently than greater reasons to do so: Whereas greater reasons favor an alternative to some option by *outweighing* the reasons in favor of that option, exclusionary reasons undercut the putative reasons in favor of that option directly, making them in fact *no reason at all* to choose that option (see, e.g., Horty 2012; Pollock 1987; and Raz 1975).

Let me illustrate with one of my early examples of nudges and its accompanying hypothesized heuristic: the Wellness Guru's cafeteria nudge and the Satisficing Heuristic. Recall that the cafeteria nudge enables the relevant HD by presenting the options in a sequence; and the agent uses the Satisficing Heuristic by examining the options in that sequence, choosing the first one that exhibits the relevant feature to a satisfactory degree. However, she might take herself to have greater reason to do otherwise—more specifically and plausibly, to examine *all* the options and choose the one that exhibits the relevant feature to the *greatest* degree. She can think this even if, as I suggested in Section 3.1, she takes there to be *some* reason to choose the first satisfactory item presented. She might think this, for example, because she suspects that the best option is considerably better than the (merely) satisfactory option, so much so that the extra effort of

36. For a similar standard in this context, Houk has suggested that being nudged is compatible with rational decision-making only when the agent can do otherwise when she has sufficient reason to do so—and that there is no reason to think being nudged prevents this (2019: 411–12). In a less normative register, Saghai has appealed to capacities of goal-conflict recognition and resolution in spelling out what it is for a nudge to be "easily resistable", which is in turn necessary for a nudge designer's influence to be substantially noncontrolling (2013b: 499–500). Regarding my strategy, though, there are some who might deny that an agent's normative point of view can play the requisite role. Some argue that whether or not an action is attributable to the agent herself is not simply a matter of whether she performed that action in light of her normative point of view (e.g., Watson 1987; Bratman 2007); they might accordingly deny that, when a particular normative point of view within an agent is a defeater for some disposition she has, such *interruptions* are attributable to her. The argument I give in Section 4 goes some length towards defusing this worry; in brief, nothing about an agent's being nudged need prevent *her* from interrupting these HDs' manifestation in action in the stronger sense some action theorists want. For them, the present proposal—that the class of defeaters that are attributable to her are fully determined by her normative point of view—can be viewed as proof of concept.

finding it will be worth it. Alternatively, she might take herself to have exclusionary reason to do otherwise—more specifically, she might take an option's being the first satisfactory option presented to be *no reason at all* to choose it. And she might think *that*, for example, because she suspects that whoever set up the sequence of events did not have her own interests in mind, and was deliberately using the sequence to prevent her from choosing a better option. The so-called "Wellness Guru", after all, might really just be a Profits and Losses Guru.

Notice that there is an empirical question here: Does an agent's taking herself to have (greater or exclusionary) reason to do otherwise actually interrupt the relevant HD, thus preventing its manifestation in nudged action? As far as I know, this issue has not been directly studied and further testing is necessary.[37] Negative results, moreover, would be illuminating. If it turned out that people's normative point of view was ineffective against nudges, the suggestion that nudging undermines self-guidance would gain considerable force (cf. Wilkinson 2013: 345–46). Of course, such results might not be universal: Some nudges might turn out to be more easily interrupted than others, in which case concerns of self-guidance would favor only implementing those nudges that agents are able to resist via their prior normative point of view (Wilkinson 2013: 351–52).

4. Toward a Distinct Manipulation Worry

The preceding discussion addressed the following challenge: Insofar as being nudged does not undermine the agent's ability to reason well about what she shall do, the heuristics she is nudged to use must be *good* rules to reason with in the situation she is in, and she must be able to be *flexible* in their application. Based on my defense that being nudged is compatible with both these requirements, moreover, we can put together three 'design specifications' for responsible nudging. First, the nudge should enable a heuristic disposition that uses a heuristic that is a *good* rule to reason with in that situation. Second, the triggering of that disposition should involve the agent's own normative point of view. And, third, her taking there to be greater or exclusionary reason to do otherwise should be able to interrupt the disposition she has that is triggered when she is nudged.

37. See Pasek et al. (2014) for some discussion of "ballot order effects", where they note that *ambivalence* might explain some of the documented effect. And while Gaudeul and Kaczmarek's (2019) experiment with default nudges in charity donations does attempt to measure participants' attitudes toward charity, the latter measurements take place only after the nudge phase of the experiment. This means that post hoc rationalization cannot be ruled out as part of the explanation of their findings regarding participants' attitudes. And, of course, with topics like donating to charity, social desirability would probably affect the reported attitudes at any phase. My thanks to an anonymous reviewer for discussion about this experiment.

Empirical questions aside, suppose nudges can be designed that meet these specifications in a variety of situations. It is important to consider what such an effective nudge designer would then be able to *do*: He would be able to nudge an agent into choosing just what the nudge designer wants her to, relying on her to reason well toward that choice.

Suppose all the details of the following story are true. A grocer is designing the layout of his store, and, being familiar with the relevant literature on consumer psychology, places snacks with high profit margins in displays directly in front of the entrance, while relegating snacks with lower profit margins to displays directly in front of the cash registers. Furthermore, this arrangement enables the disposition to decide with the Satisficing Heuristic in many of his patrons, and the fact that satisfactory snacks for the patrons are encountered early on in their visit triggers this disposition. The arrangement of the store is therefore a nudge, and these patrons are thereby nudged to choose the high-profit snacks. The Satisficing Heuristic directly involves their individual normative point of view in the way I suggested in Section 3.1, and they can be flexible with respect to using it by taking themselves to have greater or exclusionary reasons to do other than to pick the first satisfactory option. However, many patrons who are nudged in this situation do not take themselves to have such greater or exclusionary reasons, and so their disposition to decide with the Satisficing Heuristic manifests in their choosing the high-profit snacks.

But now let us make a few normative assumptions about the case. First, assume that the fact that the grocer would make more profits on these particular snacks is not a reason for his patrons to choose them. And, second, assume that the higher-profit snacks encountered sooner are also lower quality (in terms of taste or healthiness, say) compared to the lower-profit snacks encountered later. Third, assume that the options do not differ appreciably in terms of price. Under these assumptions, it seems right that there is, in some sense, greater reason for his patrons to go for the snacks that will in fact net him lower profits.

On the picture I gave in Section 3, the patrons who perform the nudged action can be reasoning well even though they are nudged to do what they ultimately do. For that reason, their being nudged did not undermine their self-guidance, for all that has been said. However, under the normative assumptions I just laid out, such patrons will nevertheless be choosing an option (the higher-profit, lower-quality snacks encountered earlier) when there is, in some sense, greater reason for them to choose the alternative (the lower-profit, higher-quality snacks encountered later).[38] Moreover, what explains their doing this is precisely that the grocer designed and implemented a nudge to get them to do this.

38. The "in some sense" qualification here flags the fact that the layout of the store—the nudge itself—alters the landscape of normative reasons in complicating ways. We could imagine, for example, that the higher-quality snacks were not *so* much better that it would be worthwhile to

This treatment is manipulative, and plausibly objectionably so. And the best way to see the force of the objection, I think, is to say that the grocer is exerting substantial control over his patrons that is pro tanto wrong. However, if the arguments of Section 3 are on the right track, this objectionable manipulation is *not* to be explained in terms of its undermining the patrons' capacity to reason well. Indeed, it appears the grocer is *relying* on their capacity to reason well, and on that capacity operating normally, to manipulate them: He relies on their assessing whether the high-profit snacks are satisfactory, and on their not interrupting the choice they are disposed to make.[39]

In light of the fact that the grocer is exerting substantial control over his patrons, we might want to say that the grocer nevertheless undermines his patrons' autonomy. There are several ways to make this more precise that are consistent with my arguments in Section 3. First, return to the sketch of self-guidance that I provided in Section 2. There, I said a person is self-guiding only to the extent that she is a unified agent; and that she is a unified agent only to the extent that she reasons well. Now, since these are only relations of necessity (see footnote 20), this sketch allows there to be something more to unification than good reasoning, or something more to self-guidance than unification via good reasoning. So the first option is that the grocer does not undermine his patrons' capacity to reason well, but nevertheless undermines their self-guidance in some other way. Second, it is plausible that self-guidance is merely an *aspect* of autonomy, which is a multifaceted phenomenon encompassing other qualities such as self-definition, praise- and blame-worthiness, answerability, and freedom from paternalism, oppression, or indoctrination.[40] So the second option is that the gro-

double back to return the lower-quality snack once the better ones had been encountered (with the store laid out as it is). In such a situation, the transaction costs associated with the higher-quality snacks are high enough that it appears the patrons have decisive reasons to simply choose the lower-quality snacks if they already have them in hand. At the same time, since the lower-quality snacks are nonetheless satisfactory, the patrons seem to have sufficient reasons to pick them up when they are encountered. So, putting these points together, the nudge might make the patrons have sufficient reasons to pick up the lower-quality snacks, and decisive reasons to keep them once they are in hand. So when I say that, in some sense, there is greater reason for them to do otherwise, it remains to be worked out exactly what is meant.

39. This case is similar in some important respects to those discussed by Gorin, such as *Election* (see 2014a: 52–53), and Buss's (2005) discussion of Kierkegaard's Johannes and Cordelia. This is not manipulative according to how, for example, Blumenthal-Barby (2012), Blumenthal-Barby and Burroughs (2012), Hill (1991), Raz (1988), Wood (2014), and, perhaps, Wilkinson (2013) conceive of manipulation. This is because the kind of manipulation at hand precisely does *not* bypass or impede the rational decision-making processes of nudged agents, insofar as 'rational decision-making processes' refers to their reasoning well about what to do and choosing on that basis. Alternative accounts of manipulation are more amenable to my concerns, such as Gorin (2014b), Klenk (2022), and Noggle (1996).

40. For some helpful discussion on this front, though principally in a conceptual register, consult Arpaly (2003: ch. 4) and Dworkin (1993/2017).

cer leaves his patrons' self-guidance intact, but undermines their autonomy in some other respect.

It should be clear that either option exploits our contested concepts of self-guidance and autonomy, and will be more or less attractive against one's background conceptions of both.[41] So let me briefly argue for taking the second tack by elimination, drawing on a fairly common conception of self-guidance to do so.

The first option involves identifying missing elements from my sketch of self-guidance, which render the grocery-store patrons' nudged actions less than fully self-guided. A variety of theorists purport to identify these missing elements, some of which are psycho-structural in character, and some of which are historical. Despite their important differences, these theorists commonly maintain that the agent's take on her normative reasons at the time of action does not fully determine whether the motivation behind the action she ultimately takes is (authentically, genuinely) *hers*, which is a necessary condition on self-guidance vis-à-vis that motivation and that ensuing action. So, whatever else is required for such authentic motivation might be the missing element(s).[42]

What sort of missing elements might these be? Several candidates have been offered: a second-order conative endorsement of the motivation in question, where the agent is satisfied with this second-order attitude (for example, Bratman 2007: especially essays 2 and 10; and Frankfurt 1999: essay 8); the motivation being ultimately grounded in what the agent cares about (for example, Jaworksa 2007; Seidman 2009; and Shoemaker 2003); or the fact that the agent herself did not, or would not, resist the motivation's formation upon rational self-reflection (for example, Christman 1991; 1993); just to name a few.

To explain why and how the grocer undermines his patrons' self-guidance (assuming he does for the moment), we might try to point to some such missing

41. For example, Christman is likely to reject the second option because, in his view, self-guidance (or self-governance) is the genuine core of autonomy, not merely one aspect among many (Christman 2020: §1.2). And while Mele is an historicist like Christman about *autonomy*, he sharply distinguishes it from (ideal) *self-control,* where "self-control" in his sense can be construed as self-guidance (see especially the discussion at Mele 1995: ch. 7). If this construal is correct, the second option is more in the spirit of Mele's endeavor.

42. I will speak of "authentic motivation" here even though many of the theorists in this domain prefer other locutions, such as "internal motivation", "motivation with which the agent identifies", or "motivation that speaks for the agent". Prominent examples of structuralist theories can be found in Bratman (2007), Frankfurt (1988; 1999), and Jaworska (2007); and a prominent example of an historicist theory is in Christman (1991; 1993). Of course, theories can be constructed that have both structural and historical features, such as Christman's, and it is a subtle exegetical question (but beyond the scope of this paper) whether Bratman's view is not also historicist in important respects. Lastly, some theorists would resist this line of thought entirely, and say that the agent's take on her normative reasons at the time of action *does* fully determine "where she stands" vis-à-vis what to do, such that its playing its characteristic role in downstream action amounts to her self-guidance. For example, see Wallace (2006; 2014) and Watson (1975; though cf. 1987).

element. Perhaps the patrons lack a second-order conative endorsement of the right sort, or their motivation for taking the satisfactory snack is not ultimately grounded in what they care about, or they would resist the formation of that motivation upon rational self-reflection. But this approach must stipulate more than the case provides: There is no reason to think that the patrons' motivation for taking the snack encountered first is *in*authentic in one of these senses. Accordingly, there is no reason to think that the patrons fail to guide their own conduct as they take the snack from the shelf.

The more serious problem is that the grocer's manipulative strategy will be *more* effective when, and because, this motivation *is* authentic in one of these senses. After all, if the grocer's patrons *do* relate in one of these ways to their motivation to take the satisfactory snack, they will be less likely to interrupt their disposition to decide with the Satisficing Rule as it manifests in the decision to take the snack from the shelf.

One might say that this shows that the conditions for authentic motivation have been misspecified. But this reply misses the forest for the trees. In the case at hand, we find the grocer's manipulation of his patrons objectionable because of the substantial control he exerts over them, enabled by his knowledge of psychology. This assessment does not hang even implicitly on the question of whether his patrons' effective motivations are authentic or not, *whatever* the conditions for authenticity turn out to be. What's so pernicious is that the grocer's strategy works better when, and because, the agent simply has no interest in interrupting the motivations within her that nudging exploits. However one might ultimately specify the conditions for authentic motivation, it is hard to see how the grocer's strategy will not be served by his patrons' motivations being authentic, because it is *precisely* those motivations that they will have no interest in interrupting (absent conflicting motivations or normative judgments). Indeed, just as the grocer appears to be relying on his patrons' reasoning well, he can rely on his patrons' *self-guidance*.

That leaves the second option: to say that the grocer leaves his patrons' self-guidance intact, and undermines some other aspect(s) of their autonomy. Indeed, if the argument I have just given is on the right track, we can see that it is possible to *exploit* someone's self-guidance in order to undermine their autonomy in other ways. This is particularly pernicious manipulation that warrants careful consideration in future work on the ethics of nudging.

5. Conclusion

Let me close by reiterating the principal virtues of this paper.

First, my analysis of nudging allows us to see that nudging has three intimately related parts: the nudge itself, the agent being nudged, and the agent performing a nudged action. The parts, respectively, are the enabling, triggering, and manifesting of dispositions to decide with heuristics. One reason this analysis is important is that it helps us understanding just what nudging is, as free of normative baggage as possible, before we turn to ethical questions about it.

Second, this analysis lets us state more clearly one of the main ethical concerns about nudging's detrimental effects on autonomy: that being nudged undermines an agent's ability to guide her own actions, which is a key aspect of autonomy, by preventing her from reasoning well about what she shall do. Furthermore, when one person undermines another's self-guidance via nudging in this way, the former exerts a kind of substantial control over the latter that seems pro tanto wrong.

Third, however, being nudged is compatible with reasoning well after all. This is because reasoning with heuristics can still exhibit two characteristic features of reasoning well. For one thing, the heuristics can be *good* rules to reason with, provided the situation has certain familiar features such as complexity, time pressure, or implicit recommendations from more expert sources. For another, the agent herself can still be *flexible* with respect to these rules for all that has been said.

Fourth, in showing that being nudged need not prevent reasoning well, I identified three 'design specifications' that nudge design should meet: The nudge should enable dispositions that use heuristics that are *good* rules to reason with in the situation; the triggering of that disposition should involve the agent's own normative point of view; and the agent's broader normative point of view, specifically her taking herself to have greater or exclusionary reasons to do other than what the disposition disposes her to do, should be able to interrupt the manifestation of the relevant disposition.

Fifth and finally, this treatment of reasoning well while being nudged crystallizes a distinct *manipulation* worry about nudging, specifically, that it can be a way of influencing someone that does not undermine their self-guidance, but harnesses it toward the nudge designer's own ends.

Acknowledgements

The core of this paper was developed at the Cologne Institute for Economic Research, where I benefited especially from conversations with Dominik Enste and Theresa Eyerund. I am also grateful to Michael Bratman, Barry Maguire, participants of the 10th European Congress of Analytic Philosophy at Universiteit Utrecht, and anonymous reviewers for very helpful comments and discussion.

References

Alvarez, R. M., B. Sinclair, and R. L. Hasen (2006). How Much is Enough? The 'Ballot Order Effect' and the Use of Social Science Research in Election Law Disputes. *Election Law Journal*, *5*(1), 40–56.

Arpaly, Nomy (2003). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford University Press.

Blumenthal-Barby, J. S. (2012). Between Reason and Coercion: Ethically Permissible Influence in Health Care and Health Policy Contexts. *Kennedy Institute of Ethics Journal*, *22*(4), 345–66.

Blumenthal-Barby, J. S. and Hadley Burroughs (2012). Seeking Better Health Care Outcomes: The Ethics of Using the 'Nudge'. *The American Journal of Bioethics*, *12*(2), 1–10.

Boghossian, Paul (2014). What is Inference? *Philosophical Studies*, *169*, 1–18.

Bovens, Luc (2009). The Ethics of *Nudge*. In T. Grüne-Yanoff and S. O. Hansson (Eds.), *Preference Change: Approaches from Philosophy, Economics and Psychology* (207–19). Springer.

Bratman, Michael E. (2007). *Structures of Agency: Essays*. Oxford University Press.

Broome, John. (2013). *Rationality Through Reasoning*. Wiley-Blackwell.

Buss, Sarah (2005). Valuing Autonomy and Respecting Persons: Manipulation, Seduction, and the Basis of Moral Constraints. *Ethics*, *115*(2), 195–235.

Christman, John (1991). Autonomy and Personal History. *Canadian Journal of Philosophy*, *21*(1), 1–24.

Christman, John (1993). Defending Historical Autonomy: A Reply to Professor Mele. *Canadian Journal of Philosophy*, *23*(2), 281–90.

Christman, John (2020). Autonomy in Moral and Political Philosophy. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.). Retrieved from https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/

Doris, John M. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford University Press.

Doris, John M. (2018). Précis of Talking to Our Selves: Reflection, Ignorance, and Agency. *Behavioral and Brain Sciences*, *41*, 1–12.

Dworkin, Gerald (2017). Autonomy. In Robert E. Goodin, Philip Pettit and Thomas Pogge (Eds.), *A Companion to Contemporary Political Philosophy* (2nd ed., 443–51). Blackwell Publishing. (1st ed. published 1993)

Einhorn, H. J. and R. M. Hogarth (1975). Unit Weighting Schemes for Decision Making. *Organizational Behavior and Human Performance*, *13*, 171–92.

Engelen, Bart (2019). Nudging and Rationality: What Is There to Worry? *Rationality and Society*, *31*(2), 204–32.

Engelen, Bart and Thomas Nys (2020). Nudging and Autonomy: Analyzing and Alleviating the Worries. *Review of Philosophy and Psychology*, *11*, 137–56.

Frankfurt, Harry (1988). *The Importance of What We Care About*. Cambridge University Press.

Frankfurt, Harry (1999). *Necessity, Volition, and Love*. Cambridge University Press.

Gaudeul, Alexia and Magdalena C. Kaczmarek (2019). Going Along with the Default Does Not Mean Going On with It: Attrition in a Charitable Giving Experiment. *Behavioural Public Policy*. Advance online publication. https://doi.org/10.1017/bpp.2019.3

Gigerenzer, Gerd (2015). On the Supposed Evidence for Libertarian Paternalism. *Review of Philosophy and Psychology*, *6*(3), 361–83.

Gigerenzer, Gerd and Reinhard Selten (Eds.) (2001). *Bounded Rationality: The Adaptive Toolbox*. The MIT Press.

Gigerenzer, Gerd and Wolfgang Gaissmaier (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*, 451–82.

Goldstein, Daniel G., Gerd Gigerenzer, Robin M. Hogarth, Alex Kacelnik, Yaakov Kareev, Gary Klein, Laura Martignon, John W. Payne, and Karl H. Schlag (2001). Group Report: Why and When Do Simple Heuristics Work? In Gerd Gigerenzer and Reinhart Selten (Eds.), *Bounded Rationality: The Adaptive Toolbox* (173–90). The MIT Press.

Gorin, Moti (2014a). Do Manipulators Always Threaten Rationality? *American Philosophical Quarterly*, *51*(1), 51–61.

Gorin, Moti (2014b). Towards a Theory of Interpersonal Manipulation. In Christian Coons and Michael Weber (Eds.), *Manipulation: Theory and Practice* (73–97). Oxford University Press.

Grüne-Yanoff, Till (2018). Boosts vs. Nudges from a Welfarist Perspective. *Revue d'économie politique*, *128*(2), 209–24.

Grüne-Yanoff, Till and Ralph Hertwig (2016). Nudge versus Boost: How Coherent Are Policy and Theory? *Minds & Machines*, *26*, 149–83.

Hausman, Daniel M. and Brynn Welch (2010). Debate: To Nudge or Not to Nudge. *The Journal of Political Philosophy*, *18*(1), 123–36.

Hertwig, Ralph and Till Grüne-Yanoff (2017). Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science*, *12*(6), 973–86.

Hill, Thomas E. (1991). *Autonomy and Self-Respect*. Cambridge University Press.

Ho, D. E. and K. Imai (2006). Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election. *Journal of the American Statistical Association*, *101*(475), 888–900.

Ho, D. E. and K. Imai (2008). Estimating Casual Effects of Ballot Order from a Randomized Natural Experiment: California Alphabet Lottery, 1978–2002. *Public Opinion Quarterly*, *72*(2), 216–40.

Horty, John F. (2012). *Reasons as Defaults*. Oxford University Press.

Houk, Timothy (2019). On Nudging's Supposed Threat to Rational Decision-Making. *Journal of Medicine and Philosophy*, *44*, 403–22.

House of Lords, Science and Technology Select Committee (2011). 2nd Report of Session 2010–12: Behaviour Change. Authority of the House of Lords.

Hutchinson, John M. C. and Gerd Gigerenzer (2005). Simple Heuristics and Rules of Thumb: Where Psychologists and Behavioural Biologists Might Meet. *Behavioural Processes*, *69*, 97–124.

Jachimowicz, Jon M., Shannon Duncan, Ekle U. Weber, and Eric J. Johnson (2019). When and Why Defaults Influence Decisions: A Meta-Analysis of Default Effects. *Behavioural Public Policy*, *3*(2), 159–86.

Jaworska, Agnieszka (2007). Caring and Internality. *Philosophy and Phenomenological Research*, *74*(3), 529–68.

Johnson, E. J. and D. Goldstein (2003). Do Defaults Save Lives? *Science*, *302*(5649), 1338–39.

Kahneman, Daniel (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, Daniel, Paul Slovic, and Amos Tversky (Eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Klenk, Michael (2022). (Online) Manipulation: Sometimes Hidden, Always Careless. *Review of Social Economy*, *80*(1), 85–105.

Koppell, J. G. and J. A. Steen (2004). The Effects of Ballot Position on Election Outcomes. *Journal of Politics*, *66*(1), 267–81.

Krosnick, J. A., J. M. Miller, and M. P. Tichy (2004). An Unrecognized Need for Ballot Reform: Effects of Candidate Name Order. In A. N. Crigler, M. R. Just, and E. J. McCaffery (Eds.), *Rethinking the Vote: The Politics and Prospects of American Election Reform* (51–74). Oxford University Press.

Levy, Neil (2019). Nudge, Nudge, Wink, Wink: Nudging is Giving Reasons. *Ergo*, *6*(10), 281–302.

Li, D., Z. Hawley, and K. Schnier (2013). Increasing Organ Donation via Changes in the Default Choice or Allocation Rule. *Journal of Health Economics*, *32*(6), 1117–29.

Lutz, Georg (2010). First Come, First Served: The Effect of Ballot Position on Electoral Success in Open Ballot PR Elections. *Representation*, *46*(2), 167–81.

Marcinkiewicz, K. (2014). Electoral Contexts that Assist Voter Coordination: Ballot Position Effects in Poland. *Electoral Studies*, *33*, 322–34.

McHugh, Conor and Jonathan Way (2018a). What is Good Reasoning? *Philosophy and Phenomenological Research*, *96*(1), 153–74.

McHugh, Conor and Jonathan Way (2018b). What is Reasoning? *Mind*, *127*(505), 167–96.

Mele, Alfred (1995). *Autonomous Agents: From Self-Control to Autonomy*. Oxford University Press.

Meredith, M. and Y. Salant (2013). On the Causes and Consequences of Ballot Order Effects. *Political Behavior*, *35*(1), 175–97.

Miller, Joanne and Jon A. Krosnick (1998). The Impact of Candidate Name Order on Election Outcomes. *Public Opinion Quarterly*, *62*(3), 291–330.

Morton, Jennifer M. (2011). Toward an Ecological Theory of the Norms of Practical Deliberation. *European Journal of Philosophy*, *19*, 561–84.

Morton, Jennifer M. (2017). Reasoning under Scarcity. *Australasian Journal of Philosophy*, *95*, 543–59.

Noggle, Robert (1996). Manipulative Actions: A Conceptual and Moral Analysis. *American Philosophical Quarterly*, *33*(1), 43–55.

Pasek, Josh, Daniel Schneider, Jon A. Krosnick, Alexander Tahk, Eyal Ophir, and Claire Milligan (2014). Prevalence and Moderators of the Candidate Name-Order Effect: Evidence from Statewide General Elections in California. *Public Opinion Quarterly*, *78*(2), 416–39.

Payne, J. W., J. R. Bettman, and E. J. Johnson (1993). *The Adaptive Decision Maker*. Cambridge University Press.

Pollock, John L. (1987). Defeasible Reasoning. *Cognitive Science*, *11*, 481–518.

Raz, Joseph (1975). *Practical Reasoning and Norms*. Hutchinson and Company.

Raz, Joseph (1988). *The Morality of Freedom*. Clarendon Press.

Rozeboom, Grant J. (2020). Nudging for Rationality and Self-Governance. *Ethics*, *131*, 107–21.

Saghai, Yashar (2013a). Salvaging the Concept of Nudge. *Journal of Medical Ethics*, *39*(8), 487–93.

Saghai, Yashar (2013b). The Concept of Nudge and its Moral Significance: A Reply to Ashcroft, Bovens, Dworkin, Welch and Wertheimer. *Journal of Medical Ethics*, *39*(8), 499–501.

Samuelson, William and Richard K. Zeckhauser (1988). Status Quo Bias in Decision Making. *Journal of Risk and Incertainty*, *1*, 7–59.
Schmidt, Andreas T. (2019). Getting Real on Rationality—Behavioral Science, Nudging, and Public Policy. *Ethics*, *129*(4), 511–43.
Schmidt, Andreas T. and Bart Engelen (2020). The Ethics of Nudging: An Overview. *Philosophy Compass*, *15*(4), 1–13.
Seidman, Jeffrey (2009). Valuing and Caring. *Theoria*, *75*(4), 272–303.
Selinger, Evan and Kyle Whyte (2011). Is There a Right Way to Nudge? The Practice and Ethics of Choice Architecture. *Sociology Compass*, *5*(10), 923–35.
Shoemaker, David W. (2003). Caring, Identification, and Agency. *Ethics*, *114*, 88–118.
Simon, Herbert (1956). Rational Choice and the Structure of the Environment. *Psychological Review*, *63*(2), 129–38.
Simon, Herbert (1957). *Models of Man*. John Wiley.
Sunstein, Cass R. (2015). The Ethics of Nudging. *Yale Journal on Regulation*, *32*(2), 413–50.
Thaler, Richard H. and Cass R. Sunstein (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.
Tversky, Amos and Daniel Kahneman (1973). Availability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology*, *5*(2), 207–32.
Tversky, Amos and Daniel Kahneman (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–31.
van Dalen, H. P. and K. Henkins (2014). Comparing the Effects of Defaults in Organ Donation Systems. *Social Science and Medicine*, *106*, 137–42.
Wallace, R. Jay (2006). Caring, Reflexivity, and the Structure of Volition. In *Normativity and the Will* (190–211). Clarendon Press.
Wallace, R. Jay. (2014). Reasons, Policies, and the Real Self: Bratman on Identification. In Manuel Vargas and Gideon Yaffe (Eds.), *Rational and Social Agency: The Philosophy of Michael Bratman* (106–28). Oxford University Press.
Watson, Gary (1975). Free Agency. *The Journal of Philosophy*, *72*(8), 205–20.
Watson, Gary (1987). Free Action and Free Will. *Mind*, *96*, 154–72.
Webber, R., C. Rallings, G. Borisyuk, and M. Thrasher (2014). Ballot Order Positional Effects in British Local Elections, 1973–2011. *Parliamentary Affairs*, *67*(1), 119–36.
Wedgwood, Ralph (2006). The Normative Force of Reasoning. *Noûs*, *40*(4), 660–86.
Wedgwood, Ralph (2012). Justified Inference. *Synthese*, *189*(2), 1–23.
Wilkinson, T. M. (2013). Nudging and Manipulation. *Political Studies*, *61*(2), 341–55.
Wood, Allen W. (2014). Coercion, Manipulation, Exploitation. In Christian Coons and Michael Weber (Eds.), *Manipulation: Theory and Practice* (17–50). Oxford University Press.