If Nudges Treat Their Targets as Rational Agents, Nonconsensual Neurointerventions Can Too

THOMAS DOUGLAS

Oxford Uehiro Centre for Practical Ethics
Faculty of Philosophy
University of Oxford

[This is a pre-publication version of an article due to appear in *Ethical Theory & Moral Practice*: https://www.springer.com/journal/10677]

Abstract. Andreas Schmidt and Neil Levy have recently defended nudging against the objection that nudges fail to treat nudgees as rational agents. Schmidt rejects two theses that have been taken to support the objection: that nudges harness irrational processes in the nudgee, and that they subvert the nudgee's rationality. Levy rejects a third thesis that may support the objection: that nudges fail to give reasons. I argue that these defences can be extrapolated from nudges to some nonconsensual neurointerventions; if Schmidt's and Levy's defences succeed, then some nonconsensual neurointerventions neither harness irrationality, nor subvert rationality, nor fail to give reasons. This, I claim, poses a challenge both to opponents of nonconsensual neurointerventions, and to defenders of nudging.

Keywords: nudging, manipulation, neurointerventions, rational persuasion, giving reasons, bypassing.

1. INTRODUCTION

We often decide how to act not through slow, careful and conscious deliberation, but by employing simple, quick and sometimes subconscious heuristics such as 'choose what is most salient', 'stick with the default' or 'listen to people you recognise'. Heuristic-triggering nudges—henceforth simply 'nudges'—influence our decisions by arranging our environment so as to prompt the use of such

heuristics.[1] These interventions received their canonical introduction and defence in Richard Thaler and Cass Sunstein's 2008 book *Nudge*, and have since been widely deployed by policymakers. They have also become the target of numerous ethical objections. One such objection holds that nudges fail to treat nudgees as befits their rational agency or, as I will henceforth paraphrase this, that they fail to treat nudgees *as rational agents*, or simply *as rational*.

In recent works, Andreas Schmidt (2019) and Neil Levy (2017; 2018; 2019) have defended nudging against this objection, in each case, by rejecting theses that have been, or might be, taken to support it. In what follows, I will argue that these defences can be extrapolated from nudges to some *nonconsensual neurointerventions*. These are interventions that alter a person's neural states through 'direct' means—means other than engaging perceptual processing—and that are performed without the target's consent (Pugh and Douglas 2017). Widely discussed examples include the administration, as part of a criminal sentence, of drugs to suppress sexual or addictive desires in persons convicted of sexual or drug-related crimes.[2] Less widely discussed but more widely employed examples include the mandatory administration of anti-psychotics or anti-depressants to psychiatric patients deemed to pose a risk to themselves or others.

That Schmidt's and Levy's defences of nudging can be extended to some nonconsensual neurointerventions is, I claim, an interesting result, for these interventions are highly controversial—much more so than nudges.[3] Indeed, even their staunchest defenders advocate their use only in very specific contexts and under stringent safeguards.[4] Moreover, the preeminent criticism of nonconsensual neurointerventions has been precisely that they fail to treat their targets as rational. Jan Christoph Bublitz holds that such neurointerventions are 'objectifying and disrespectful of the targeted person as a rational and self-

[1] Many would favour a broader conception of nudges on which some nudges do not prompt the use of heuristics. Barton and Grüne-Yanoff (2015: 343, their italics) distinguish nudges which '*trigger* the use of certain heuristics' from nudges that 'counteract or *block* the detrimental use of heuristics in certain environments' (such as cooling-off periods) and nudges that 'have no special connection with heuristics at all' (such as the provision of information about dangers). (See also Engelen 2019: 219.) I limit my focus to heuristic-triggering nudges because these are the interventions that critics typically have in mind when they advance the critique that is my focus in this article—that nudges fail to treat nudgees as rational agents. Limiting my focus in this way is thus more charitable to those critics than adopting a broader, albeit arguably more standard, definition.

[2] Such interventions are permitted in a number of US states and continental European jurisdictions. See, for discussion, Forsberg (2021).

[3] For a review of the debate, see Pugh and Douglas (2017). For a collection containing many of the subsequent contributions to it, see Birks and Douglas (2018).

[4] For a comprehensive but qualified defence of nonconensual neurointerventions in the context of criminal justice, see Ryberg (2019).

controlling' (2018: 303); Elizabeth Shaw wonders whether nonconsensual neurointerventions 'could be viewed as objectification, treating the individual merely as a means, or failing to respect the individual's rational agency' (2018: 3); and Christopher Bennett suggests that they violate the requirement that '[w]e treat the person as an equal by dealing with them in the image of a rational agent' (2018: 265).[5] My argument poses a challenge to these views. It also, I will suggest, poses a challenge to defenders of nudging, since it raises the prospect that their defences show too much, establishing that some nonconsensual neurointerventions treat their targets as rational when, intuitively, they do not.

My argument proceeds as follows. I begin, in §2, by introducing two theses that have been advanced in support of the view that nudges fail to treat nudgees as rational: that they harness irrationality, and that they subvert rationality. I then, in §3, outline Schmidt's defence of nudging, which consists in a rejection of these two theses, before arguing, in §4, that this defence can also be invoked against the analogous theses regarding nonconsensual neurointerventions: Schmidt's defence implies that some nonconsensual neurointerventions neither harness irrationality nor subvert rationality. In §5, I consider how Schmidt might nevertheless accommodate the intuitively plausible view that nonconsensual neurointerventions invariably fail to treat their targets as rational. I land on the view that he might do so by maintaining that, even when they neither harness irrationality nor subvert rationality, nonconsensual neurointerventions fail to 'give reasons', and thereby fail to treat their targets as rational. However, I note that this response calls into question whether paradigmatic nudges really treat their targets as rational, for it is not obvious that these nudges 'give reasons' either.

This brings us to Levy's defence of nudging. Levy has recently argued that nudges typically do give reasons. I introduce this defence in §6. However, I then argue, in §§7-9 that, on its most plausible understanding, Levy's defence—like Schmidt's—extends also to some nonconsensual neurointerventions; it implies that some nonconsensual neurointerventions also give reasons, though, intuitively, they do not treat their targets as rational. I end, in §10, by drawing out some of the further normative implications of my argument, explaining why it poses a challenge to both opponents of neurointerventions and defenders of nudging.

[5] For critical discussion of the view that nonconsensual neurointerventions fail to treat their targets as rational, see Ryberg (2018: 187-8; 2019: 115ff, 129ff).

2. SUBVERTING RATIONALITY AND HARNESSING IRRATIONALITY

Let me begin, then, by distinguishing two theses that have been advanced in support of the view that nudges fail to treat nudgees as rational agents. [6]

The first has been characterised variously as holding that nudging 'perverts decision-making' (Wilkinson 2013: 349), 'encourage[s]' or 'foster[s]' irrationality (Cohen 2013: 5), 'perverts people's rationality and thus makes them less rational' (Engelen 2019: 206), or 'undercuts people's rational agency' (Schmidt 2019: 515; see also Bovens 2009). I will call this view the *subverting rationality thesis* and will take it to hold that nudging diminishes the nudgee's procedural rationality (Engelen 2019: 220); it results in processes that are less rational than the processes that would otherwise have obtained.[7]

According to the second thesis, the problem is not that nudges *diminish* the nudgee's procedural rationality, compared to the situation that would have obtained had the nudge not been employed, but simply that they operate *via* irrational processes. Luc Bovens suggests that it may be characteristic of nudging 'that some pattern of irrationality is being exploited' (2009: 209); Sarah Conly holds that, as nudgers, '[r]ather than regarding people as generally capable of making good choices, we outmaneuver them by appealing to their irrationality' (2013: 30); and Schmidt assesses the view that nudges 'work through psychological mechanisms that deviate from traditional notions of rationality' (2019: 511). I will call this view the *harnessing irrationality thesis*, and will take it to hold that nudging affects the nudgee's decisions at least in part via an irrational process in the nudgee.[8]

---

[6] For discussion of similar theses regarding manipulation—a category of influence that overlaps substantially with (and arguably subsumes) the category of nudging—see Gorin (2014a, 2014b).

[7] I interpret this thesis globally, as referring to the total set of processes that will occur in the nudgee, but one can also imagine local variants of it. These would focus on some subset of that set, for example, those processes that are either temporally or causally proximate to the nudgee's decision. Moving to a local interpretation of the thesis would not affect the substance of my argument below, though it would necessitate some small presentational changes. I adopt the global variant for ease of explication.

[8] I borrow the term 'harnessing' from Schmidt, who at one point also characterises the thesis as being that nudges 'harness systematic irrationality' (2019: 542). I prefer the term 'harnessing' to the oft-used 'exploiting', since it is less normatively loaded. As with the subverting rationality thesis, I interpret this thesis globally, as referring to all processes in the nudgee that are causal intermediaries between the nudge and the nudged decision. That is, I take the claim to be that this whole process, or at least one subprocess thereof, is irrational. Local variants would focus solely on some subset of these processes—perhaps those that are causally or temporally proximate to the final decision, and I will have cause to consider one such local variant below. However, elsewhere I adopt the global variant, for ease of explication.

The subverting rationality thesis and the harnessing irrationality thesis are not always distinguished from one another. But they are distinct, and moreover, they can come apart. A nudge could exert its effects on decisions (in part) via processes that are irrational, but no more so than the processes that would otherwise have occured.[9] For example, the nudge may simply take advantage of an irrational process that would have occurred—and been equally irrational—regardless; the nudge may do no more than arranging the choice context such that this process favours one choice rather than another. This nudge would harness irrationality, but without subverting rationality.

It is also conceivable that a nudge could diminish the procedural rationality of the nudgee (thus subverting rationality) but without operating via any irrational process (thus not harnessing irrationality). Perhaps the nudge subverts the nudgee's rationality only by diminishing the rationality of processes that are not *harnessed* by the nudge—*viz.*, that do not mediate the nudge's effects on the agent's decisions—but are merely side-effects of it. Or perhaps the nudge diminishes the nudgee's procedural rationality, but not to the degree that the nudgee's processes count as *irrational* (as opposed to merely 'arational' or 'less-than-fully rational').[10]

3. SCHMIDT'S DEFENCE

In his recent (2019) defence of nudging, Andreas Schmidt seeks to refute the harnessing irrationality and subverting rationality theses, and thereby to undermine the objection that '[p]ublic policy nudging implies treating agents as irrational' (p. 516).[11] His implicit assumption here is that, if a nudge neither harnesses irrationality nor subverts rationality, then it treats its targets as rational. Or at least, there will be no good reason to deny this. However, for the moment, I will limit myself to exploring Schmidt's arguments against the

[9] Variants of this point have been made by numerous contributors to the ethical debate on nudging. See, for example, Thaler and Sunstein (2003: 175; 2008: 2-4, 243, 247); Anderson (2010: 372-3); Hausman and Welch (2010: 132-3); Grill (2014); Barton and Grüne-Yanoff (2015: 348); Wilkinson (2017); Engelen (2019); and Schmidt (2019: 530-1). Note that, if the subverting rationality thesis is interpreted globally—as maintaining that the nudgee's total future decision-making processes will, all things considered, be less rational than had the nudge not occurred—then there is also the possibility that an irrationality-harnessing nudge could avoid subverting rationality because, though it diminishes rationality in the short term, it has a longer-term rationality-enhancing effect.

[10] I will remain neutral on whether, to qualify as rational, a process must (a) be *fully* rational or (b) simply exceed some threshold level of rationality. I will also remain neutral on whether, (c) necessarily, processes that are not rational are irrational, or (d) processes that are not rational can be either irrational or *a*rational.

[11] Schmidt does not explicitly present the subverting rationality thesis, but, as we will see, he does advance a claim that contradicts it.

harnessing irrationality and subverting rationality theses, setting aside the matter of what their falsity further implies.

Schmidt begins his critique of both theses by defending an ecological conception of rationality, according to which 'a person's decision is procedurally rational in an environment to the extent that, given her particular psychological makeup, the decision-making procedures she uses allow her to reliably achieve her ends in this type of environment' (p. 521). (The 'reliably' implies that the processes would likely further the agent's ends even if the circumstances, or the agent's psychology, were changed slightly.) He argues that, on this conception, rational processes may be 'satisficing' rather than maximising, insensitive to certain information (even when that information is relevant to one's ends), and unconscious and automatic rather than conscious and deliberative (p. 522). This allows the heuristics triggered by many nudges to qualify as rational, with the upshot that these nudges do not harness irrationality—at least, not by virtue of triggering these heuristics. Schmidt concludes that 'the decision-making procedures through which nudges work need not be irrational' (p. 527)—the harnessing irrationality thesis fails.

Might nudges nevertheless invariably *subvert* the nudgees' rationality? They might. For example, perhaps the heuristics harnessed by nudges—though not irrational—are nevertheless *less* rational than those that would otherwise have obtained. However, Schmidt argues that, for many nudges, the reverse will be true; the nudge will *enhance* the nudgee's procedural rationality, on his ecological account.[12] What ecological rationality requires is 'that there is a good match between (a person's) decision-making procedures, her particular psychological makeup, and her choice environment' (p. 528). By adapting a person's choice environment to her psychological makeup, nudges can, Schmidt claims, frequently improve this match (pp. 530-1).[13] Schmidt denies, moreover, that this direct rationality-enhancing effect of nudges need be accompanied by any other erosion of procedural rationality (pp. 536-40).

Consider this paradigmatic nudge:

> *Salience Nudge.*[14] The staff in a prison cafeteria would like to encourage healthier eating, so they introduce a new policy: all and only the healthiest foods will be placed at eye level in the cafeteria refrigerators. Psychological evidence shows that foods placed at eye

[12] Similar points are made by Blumenthal-Barby (2012: 356) and Blumenthal-Barby and Naik (2015).

[13] For further recent responses to the subverting rationality claim, see Wilkinson (2017) and Engelen (2019).

[14] This case is inspired by a case given by Thaler and Sunstein (2008: 1-2). I will use the label '*Salience Nudge*' to refer to both the case and the intervention described therein, ensuring it is clear from the context which reference is intended. I employ a similar approach with all other cases to which I give italicised labels in this article.

> level are, other things being equal, more salient to customers than other foods, and that people tend to subconsciously favour more salient items when making their food selections. The prisoners are informed about the intervention, and the reasons why it is being employed.[15] However, prisoners who are unhappy about being nudged in this way have no reasonable way of avoiding it; there is no alternative source of food in the prison.[16] As a result of the new policy, some prisoners choose the healthiest foods when they would otherwise have chosen less healthy foods.

Whether the processes harnessed by this nudge reliably further the prisoners' ends in the type of environment in which they find themselves is perhaps open to question. The answer may depend on how finely we specify the 'type of environment'. Are we thinking of environments in which one much choose between alternative items? Between alternative foods? Between alternative foods in a prison cafeteria? It may also depend on how products are normally arranged. Are foods placed at eye level typically products that satisfy consumer preferences? Or are foods normally arranged simply so as to maximise their shelf-life or minimise the time taken to re-stock shelves? Thus, whether *Salience Nudge* harnesses irrationality is perhaps unclear.

Nevertheless, it is very plausible that *Salience Nudge overall increases the degree to which* those processes reliably further the nudgees' ends in the circumstances in which they find themselves, and thus enhances, rather than subverting, the prisoners' rationality. Most prisoners will presumably have ends—such as living to see their grandchildren, and being able to remain active into old age—whose realisation will be facilitated by choosing healthier foods in the prison cafeteria. And many of these prisoners will also have no other comparably important ends that will be frustrated by this choice. Choosing the healthy foods will, all things considered, further the ends of these prisoners.[17] And since *Salience Nudge* arranges the prisoners' environment in such a way that their responses to food salience more reliably produce this choice, it increases the degree to which those processes reliably further their ends. It enhances their procedural rationality, on Schmidt's ecological account of rationality.

---

[15] I make this stipulation in order to exclude transparency-based objections to the nudge, which are not my focus here.

[16] I make this stipulation in order to rule out the possibility that the prisoners can be taken to have validly consented to the nudge, which might be thought to immediately foreclose the possibility that the nudge fails to treat the prisoners as rational.

[17] I assume that Schmidt intends 'furthering' to be understood as an all-things-considered concept. Thus, it is not enough, to further an agent's ends, that something *in some respect* helps her to achieve her ends, or leads to her achieving *some* of her ends. I henceforth adopt this understanding and omit 'all things considered' when speaking of furthering ends.

4. EXTENDING SCHMIDT'S DEFENCE

Though intended only as a defence of nudging, Schmidt's defence has interesting implications for certain non-nudge forms of behavioural influence. Consider:

> *Thirst Drug.* Due to the cool temperature in the prison cafeteria, prisoners tend to drink little water. The cafeteria staff would like to encourage prisoners to drink more, in order to improve their health, so they arrange to have a thirst-enhancing drug sprayed into the air. The drug is a hormone that the body also produces naturally in response to dehydration, and that promotes thirst. The prisoners are informed about the intervention, and the reasons why it is being employed. They can also see the spray being released. However, prisoners who are unhappy about being administered the drug have no reasonable way of avoiding it; there is no alternative place to eat in the prison. As a result of the intervention, some prisoners take a cup of water when they would not otherwise have done so.

The intervention described in *Thirst Drug* is a *nonconsensual neurointervention*—an intervention that nonconsensually and intentionally alters a person's neural states other than through engaging perceptual processes (Pugh and Douglas 2017). (Though the prisoners in *Thirst Drug see* the spray, the intervention does not exert its intended effects on the prisoners' brain states *via* this perceptual process.) *Thirst Drug* is also, at least arguably, *not* a nudge, since it is plausibly definitional of nudging that it operates by arranging the nudgees' *perceived* environment. Nevertheless, the intervention in *Thirst Drug* is certainly one of which we can legitimately ask: Does it harness irrationality? And does it subvert rationality? And Schmidt's defence of nudging appears to give us the resources we need to answer these questions. Let us, then, consider what Schmidt's arguments imply for *Thirst Drug.*

Consider first the question about harnessing irrationality. The neuro-chemical regulation of thirst, and our psychological responses to thirst, plausibly qualify as rational processes, on Schmidt's ecological account, since these processes typically militate in favour of preserving a relatively healthy level of hydration, which generally furthers our ends, all things considered.[18] It is thus doubtful that *Thirst Drug* harnesses any irrational process.

Consider next the question about subverting rationality. If we assume that prisoners ends are better served by drinking water than by not doing so, then we

[18] This is true even though, as it happens, these mechanisms malfunction slightly, by failing to produce optimal levels of water consumption in the cool temperatures of the cafeteria; the neuro-chemical and psychological regulation of thirst can *somewhat reliably* further our ends, even if it does not *optimally* do so.

can expect that *Thirst Drug* will, on Schmidt's ecological conception of rationality, enhance, rather than subverting, their rationality. As with *Salience Nudge*, *Thirst Drug* can credibly be thought of as better matching these prisoners' psychology—and in this case also physiology—to their environment. It does so by arranging the environment such that the prisoners' psychological responses to thirst, and physiological processes that produce thirst, more reliably further their ends.

5. GIVING REASONS

I have been suggesting that if, for the reasons given by Schmidt, some nudges neither harness irrationality nor subvert rationality, then nor does *Thirst Drug.* This creates a problem for Schmidt, since it is intuitively plausible that *Thirst Drug does* fail to treat the prisoners as rational. We might thus wonder: how might Schmidt explain why *Thirst Drug* fails to treat its targets as rational?[19] What explains why *Thirst Drug* fails to treat its targets as rational even though, on Schmidt's understanding of these concepts, it neither subverts their rationality, nor harnesses their irrationality?

(A) Harnessing Physiological Processes. One answer would invoke the thought that only *psychological* processes are apt for classification as irrational or rational, whereas at least some of the processes harnessed by *Thirst Drug*—those that produce the feeling of thirst, as opposed to responding to it—are physiological, not psychological. That is to say, they are physico-chemical processes with no mental correlates. It might be thought that these processes are simply *arational*. Thus, it might be held, even if *Thirst Drug* does not harness irrationality, it does harness *a*rationality, and that, perhaps, is sufficient to show that it fails to treat its targets as rational.[20]

Notice, however, that similar thoughts apply to all nudges. Take *Salience Nudge. Salience Nudge* operates in part via processes—such as the transmission of photons through the eyeball—that are physiological, in the sense defined above: they have no mental correlates. And the same applies to all other nudges. Thus, if harnessing arational physiological processes is sufficient for an influence to fail to treat as rational, the project of defending nudges will be hopeless from the outset. This answer will thus not be an attractive one for Schmidt and other proponents of his defence.

More importantly, it is simply not plausible that harnessing physiological processes precludes treating as rational. Even forms of influence that are paradigmatic examples of 'treating as rational' harness such processes. Consider:

---

[19] *Thirst Drug* may also wrong the prisoners other than by failing to treat them as rational. For example, it may infringe their right to bodily integrity.

[20] I thank an anonymous reviewer for *Ethical Theory and Moral Practice* for pressing me to consider this objection.

> *Health Claims.* The staff in a prison cafeteria would like to encourage healthier eating, so they introduce a new policy: when cafeteria staff address each prisoner, they explicitly state one good reason for choosing one of the healthiest options available. They say things like 'I suggest the spinach salad; spinach contains a lot of vitamins' or 'broccoli is very good for you—it's high in antioxidants'. As a result, some prisoners choose healthier foods than they would otherwise have chosen.

This intervention surely treats its targets as rational, even if, as in *Thirst Drug*, the prisoners have no reasonable means of avoiding being exposed to it. Yet, like *Thirst Drug* and *Salience Nudge*, this intervention operates in part via physiological processes, such as the vibration of the prisoners' ear drums as they hear the health claims being uttered. If only psychological processes can be rational or irrational, this mechanical process must be an arational process. Thus, *Health Claims* appears to harness an arational process. Yet it would be implausible to claim on this basis that *Health Claims* fails to treat the prisoners as rational.

In response, Schmidt might hold that, in assessing whether an intervention harnesses arationality, we should examine only the *final stage* in the process via which the intervention exerts its effect on its target's decisions—the process via which that decision is ultimately made. On this approach, whether an intervention harnesses arationality would depend solely on whether that final stage of the process is arational. In *Health Claims*, and perhaps also *Salience Nudge*, it is plausible that this final stage in the process is rational. For example, in *Health Claims*, the prisoners presumably arrive at their decision regarding what foods to eat by *thinking about* the information that the staff have provided. By contrast, it might be held that in *Thirst Drug*, the whole process via which the intervention exerts its effect on the prisoners' decisions—including the final stage of that process—is arational. Thus, on the present view, *Thirst Drug* harnesses irrationality, but *Health Claims* does not.

The difficulty with this response is that it is simply not clear that the final stage of the process via which *Thirst Drug* exerts its effect *is* arational. Indeed, we can refine the case so as to make it clear that it is not. Suppose that, once the prisoners experience the feelings of thirst induced by the drug, they then reflect on whether, in the light of those feelings, they would like to take a cup of water. Some prisoners decide that they would, others do not. If it is sufficient, to avoid harnessing arationality, that an intervention operates via a process that ends in a rational process, we must conclude that this refined version of *Thirst Drug* does not harness arationality. Yet it remains plausible that *Thirst Drug* fails to treat the prisoners as rational.

(B) Failing to Give Reasons. There is, however, a more promising basis on which Schmidt might hold that *Thirst Drug* fails to treat the prisoners as rational. This basis can be motivated by reflecting further on *Health Claims*. A plausible explanation of why *Health Claims* treats the prisoners as rational would appeal to the thought that it involves *giving reasons*—a form of influence that is standardly taken to treat as rational (see, for example, Shiffrin 2000: 213; Blumenthal-Barby 2012: 351). The cafeteria staff induce prisoners to choose healthier foods by presenting them with *reasons* to choose those foods.

In *Health Claims*, the cafeteria staff *explicitly* give the prisoners reasons, but there might be other ways of giving reasons too. Consider:

> *Pictures.* All is as it is in *Health Claims* except that, this time, rather than explicitly stating reasons, the cafeteria staff merely draw prisoners' attention to the health consequences of their choices. Healthy foods are accompanied by a picture of a healthy-looking heart. As a result, some prisoners choose healthier foods than they would otherwise have chosen.

It is plausible to think that the staff treat the prisoners as rational, and we can again explain why by maintaining that they give the prisoners reasons, albeit this time implicitly rather than explicitly.

One further type of case should be mentioned. In both *Health Claims* and *Pictures*, the cafeteria staff give reasons by drawing attention to pre-existing reasons, but we can also give one another reasons by creating new reasons. Consider:

> *Incentives.* The staff in a prison cafeteria would like to encourage healthier eating, so they introduce a new policy: prisoners who choose the healthier foods will be allowed to use the prison's games room for longer periods than previously; prisoners who choose less healthy foods will not. As a result, some prisoners choose healthier foods than they would otherwise have chosen.

Here again, we can explain why the intervention treats the prisoners as rational—as I suppose it does—by maintaining that it involves giving prisoners a reason to choose the healthier foods.[21] By contrast, it might be thought that *Thirst Drug* does not give the prisoners reasons, whether by creating new reasons or drawing attention to pre-existing ones. Perhaps this explains why it fails to treat the

[21] I am not suggesting here that creating reasons *always* treats the reason-receiver as rational. Coercion arguably operates by creating reasons, yet it is at least open to question whether it treats the coercee as rational.

prisoners as rational, though it neither subverts rationality nor harnesses irrationality.

However, this explanation also calls into question whether paradigmatic nudges really treat nudgees as rational. Recall *Salience Nudge*, in which healthier foods were made salient simply by placing them at eye level. Does this intervention give reasons? Perhaps. Suppose that it is easier to reach for foods placed at eye level than to reach for foods placed in other locations. In placing foods at eye level, cafeteria staff will thus arguably have given the prisoners prudential reasons to choose those foods.[22] They will have given a reason by *creating* a reason. However, suppose this is not so. Suppose it is in fact somewhat easier to reach for foods placed at waist level than those placed at eye level. Nevertheless, because foods at eye level are more salient, the prisoners are more likely to choose them. In that case, would the cafeteria staff have given the prisoners a reason to choose the healthier foods? Not obviously. The cafeteria staff in this case do not explicitly state reasons (as in *Health Claims*), and it is not obvious that they draw attention to them (as in *Pictures*).[23] Indeed, nudges are standardly understood not to give reasons.[24] Thus, if we suppose that *Thirst Drug* fails to treat its targets as rational because it fails to give reasons, we are in need of a further defence of nudging. We need an account of why at least some nudges *do* give reasons.

6. LEVY'S DEFENCE

Schmidt does not offer such an account—he is interested in the nature, and especially the rationality, of the processes harnessed by nudges, not in the details of *how* those processes are harnessed (by giving reasons or otherwise). However, in his recent defences of nudging, Neil Levy does begin to develop such an account. Levy confronts the worry that nudges 'bypass our reasoning processes' (2019: 281; see also Levy 2017; 2018). He does not say exactly what 'bypassing' consists in, but his response to the worry suggests that he thinks a nudge could bypass reasoning in either of two different ways: by failing to harness any rational process

[22] I thank an anonymous reviewer for pressing me to consider this possibility.

[23] Rozeboom (2020) raises a similar worry: that nudges—even rationality-enhancing nudges—might fail to treat the nudgee as a rational agent by failing to recognise her rational authority. I take this to be one diagnosis of why, if nudges fail to give reasons, they are morally problematic. But I will not commit to this diagnosis here. I will understand the worry more generically: if nudges, such as *Salience Nudge*, fail to give reasons, then they fail to the nudgees as befits their rationality.

[24] See, for example, Blumenthal-Barby (2012: 349), who illustrates her category of 'Reason-Bypassing Nonargumentative Influence' by giving the examples of 'framing, setting up defaults, setting up the environment a certain way, and priming using subconscious cues'.

in the nudgee, or by harnessing a rational process other than by giving it a reason (henceforth just 'failing to give reasons'). Consider the following:

> Nudging doesn't bypass our capacity to reason. When they are effective in changing behavior, [nudges] typically (though perhaps not invariably) work by giving us reasons. These reasons may not be consciously recognized or responded to by agents, but they are reasons nevertheless, and it is in virtue of being reasons that they alter behavior. The mechanisms that respond to nudges are reasoning mechanisms, and in most cases, at least, nudges no more bypass reasoning than do philosophical arguments. (Levy 2019: 282-3)

Levy's claim that '[t]he mechanisms that respond to nudges are reasoning mechanisms' can be viewed as his response to the harnessing irrationality thesis, and it is to the defence of this claim that Levy devotes most of his attention.[25] But Levy also makes a further claim, which will be my focus here: he holds that nudges typically 'work by giving us reasons'. This can be viewed as his response to the worry that I raised in the previous section—the worry that nudges fail to give reasons.

Levy does not make out the case for this further claim in detail. In particular, he does not explicitly state the basis on which he takes nudges to operate by giving reasons (henceforth just 'to give reasons'). Nevertheless, he does offer some clues as to what this basis might be. In what follows, I will consider two possible bases, inspired by—though extending somewhat beyond—Levy's own discussion: one appeals to the view that nudges communicate recommendations; the other appeals to the view that they communicate reasons. I will argue that the first fails to establish that nudges typically give reasons, while the second implies that some nonconsensual neurointerventions give reasons too.

7. COMMUNICATING RECOMMENDATIONS

The exemplars that Levy invokes to support his claim that nudges typically give reasons are nudges that, like *Salience Nudge*, operate by making one option—the option favoured by the nudger—especially salient to the nudgee, for example, by placing it first on a list of options or otherwise making it more visually prominent than the alternatives. Levy suggests that these salience-based nudges are effective because salience 'is taken to be an implicit recommendation. There is evidence that people … tend to see options that have been made salient to them as having been recommended to them' (2019: 290). He continues:

> If it is rational to be guided in our judgments by testimony (and it surely is, under many conditions), it is no less rational to be guided by implicit

[25] In rejecting the harnessing irrationality thesis, Levy goes further than Schmidt, and in two different ways. First, he does not merely claim that the processes harnessed by some nudges are not irrational, he claims that they are rational processes of a particular kind: they are *reasoning* processes. Second, he claims not only that *some* nudges harness reasoning processes but that nudges *typically* harness such processes.

> recommendations … Setting defaults or framing options is communicative: people frame options in ways that highlight particular choices because they take them to be good ones, and their communicative intent is recognised by those who respond to the framing. (2019: 290)

There are two crucial claims here: that salience-based nudges typically implicitly express recommendations, and that nudgees are typically moved by those nudges in part because they *recognise* this recommendation—because the recommendation is successfully communicated to them. Levy's thought appears to be that, when they indeed work by communicating recommendations in this way, nudges give reasons. And, though he only clearly commits to the view that *salience-based* nudges typically work in this way, we might wonder whether this thought could be generalised in order to support his broader claim that nudges—salience-based or not—typically give reasons.

8. A CHALLENGE

It is, I think, plausible that nudges typically *express* recommendations. Like paradigmatic, explicit recommendations, nudges are typically motivated by the judgment that the nudgee has a reason—whether self- or other-regarding—to make the decision that the nudge is designed to promote.

Less plausible, however, is the thought that nudgees typically *recognise* those recommendations. After all, it is often not the case that those subjected to nudges *consciously* recognise that anything is being recommended. Indeed, those nudged in experimental settings typically do not consciously recognise that they are being influenced at all.[26] Rather, the thought must be that there is some kind of implicit recognition at work. But is there? As Levy notes, some empirical research does suggest that, at least in default-based nudges, nudgees implicitly *perceive* that the option chosen as the default has been recommended.[27] However, *perceiving* a recommendation is not sufficient for *recognising* it; the perception also has to be somewhat sensitive to reality—it has to track the actual presence of a recommendation. Consider, by analogy, a person who, whenever she finds a receipt lying on the ground, assumes that someone has strategically dropped the receipt as a way of implicitly recommending that she go to the shop, café, restaurant, pub, etc. that issued the receipt. This person perceives an implicit

[26] For example, Kroese et al. (2016: e135) found that only 3 out of 91 participants subjected to a food positioning nudge subsequently correctly identified the nudge.

[27] See, for example, Madrian and Shea 2001; McKenzie et al. 2006; Sher and McKenzie 2006. In what perhaps comes closest to a direct test of this, McKenzie and collaborators (2006: 417-18) presented participants with a binary choice in which either one option was presented as the default option (the test condition), or the two choices were presented in a balanced way (the control condition). Participants in the test condition were more likely to report that they had made the choice they made because it was 'what the experimenters wanted' than were participants in the control condition.

recommendation, but she does not, I think, *recognise* one, even in those rare cases where someone *has* left the receipt for precisely that reason.

In order to recognise—and not merely perceive—a recommendation, one must be sufficiently sensitive to whether something is in fact being recommended. Are nudgees typically sufficiently sensitive to this? Not obviously.

Consider *Salience Nudge*. In this case, the salience of the healthy foods arguably expresses a recommendation only because the placement of these foods at eye level was motivated by the judgment that prisoners have reasons to choose these foods. Were products distributed randomly, or, say, in such a way as to maximise their shelf-life or minimise re-stocking time, the placement of healthy foods at eye level would no longer express any recommendation. Thus, to recognise the recommendation, prisoners must be sufficiently sensitive to whether product placement was indeed motivated by a judgment about the prisoners' reasons. Yet there is no reason to suppose that the prisoners are at all sensitive to this.[28] It is plausible to suppose that, when we respond to salience-based nudges, we are responding to the salience, not the intentions for which it was created.

If salience-based nudges were *more effective* when the nudger's intentions were made clear to the nudgees, this might support the view that the nudgees are, to some degree, tracking whether the nudge expresses a recommendation. But most empirical research that has investigated this question has found that nudge-effectiveness is not enhanced by making the intentions behind the nudge clear. For example, Kroese and collaborators recently found that an experimental food-salience nudge similar to *Salience Nudge* was effective, and equally so, regardless of whether the experimenters' motives in arranging the food as they did were disclosed,[29] and similar results have been found for other salience-based nudges.[30] Thus, even if we limit ourselves to salience-based nudges, it is doubtful that

---

[28] It might be argued that placement decisions could express recommendations irrespective of the intentions of those who make them. In other contexts, we often allow that actions can have meanings that are unconnected to the intentions from which they were performed: raising my hand in the seminar room arguably expresses 'I have a question', even if I do it only to relieve an aching shoulder; turning my back on you arguably expresses 'I don't want to hear what you say', even if I do it only because I see a spider on your shoulder and am arachnophobic. In these cases, my actions seem to have the meanings that they do by virtue of the intentions that *typically* motivate actions of this kind, not by virtue of the intentions that in fact motivated them. Perhaps, then, it is enough, for *Salience Nudge* to implicitly express a recommendation, that placing a product at eye level is typically motivated by the judgment that the customers have reason to choose that product. However, this reply will not help Levy, since it is not at all clear that the prisoners in *Salience Nudge* will be sensitive to what typically motivates product placement decisions.

[29] See Kroese et al. (2016).

[30] See, for example, Steffel et al. (2016) and Bruns et al. (2018).

nudges typically communicate recommendations.[31] Levy's view that nudges typically give reasons thus remains unsupported.

9. COMMUNICATING REASONS

How might Levy respond? One option open to him would be to allow that there are other forms of communication that can give reasons besides the communication *of recommendations*. After all, in addition to expressing a proposition of the form 'I recommend that you choose the healthy foods', the placement of healthy foods at eye level in *Salience Nudge* plausibly also expresses the simpler proposition 'you have reasons to choose the healthy foods'. And it is plausible to think that expressing *this* proposition to the prisoners should count as expressing a reason. If this proposition were also *recognised* by the prisoners, perhaps Levy could claim that this is enough for the nudge to qualify as reason-giving. He might claim that one gives a reason whenever one influences by communicating a reason—whenever, that is, one influences another by expressing a proposition of the form 'you, the influencee, have reason to do x' which is then recognised by the influencee.

Do nudgees typically recognise a proposition of this form? This will depend on how sensitive nudgees typically are to whether they have the reason expressed by the nudge. In some cases, nudgees will be surely be somewhat sensitive to this. For example, some nudges may be subject to a form of rational filtering whereby those subjected to them assess—consciously or subconsciously—whether they in fact have reason to do what they are being nudged towards, and 'succumb' to the nudge only if they deem that they do. In *Salience Nudge*, for instance, the salience of the healthy foods may induce the prisoners to deliberate on which foods they have most reason to select. In that case, we would expect nudgees' responses to the nudge to be somewhat sensitive to the reasons that they in fact have. Alternatively, a nudge may work via mechanisms that track the person's reasons at a biological level. For example, perhaps *Salience Nudge* just happens to be more effective in people who have high blood sugar levels, and who thus have special reason to eat healthily. In this case too, we could aptly say that the nudgee responds to the nudge in a way that is somewhat sensitive to whether they have the reason expressed by the nudge

Do we have evidence to suggest that nudgees are *typically* sensitive to reasons in ways such as these? To my knowledge, we do not. Still, seeking such

[31] In addition to casting doubt on the suggestion that *Salience Nudge* communicates a recommendation, and thereby gives reasons, this observation may also have implications for whether *Salience Nudge* harnesses irrationality. If *Salience Nudge* is effective because the prisoners perceive a recommendation, and if that perception is unmoored from whether there is in fact a recommendation present, then the nudge arguably operates by harnessing an irrational tendency to see recommendations where there are in fact none.

evidence is, I think, the most promising route to establishing Levy's claim that nudges communicate—and thus give—reasons. And it may well turn out that this endeavour will ultimately succeed.

Note, however, that the story I've been telling in this section can—like Schmidt's defence of nudging—be extrapolated to some nonconsenual neurointerventions. On the present interpretation of Levy's defence, nudges typically give reason because it is typically the case that (a) nudges express that the nudgee has some reason, and (b) the nudgee is sufficiently sensitive, in her response to the nudge, to whether she in fact has this reason. But analogous claims will hold for some nonconsensual neurointerventions.

Recall, again, *Thirst Drug*, in which staff in the prison cafeteria spray a thirst-enhancing drug into the cafeteria air. By spraying this drug, and inducing feelings of thirst, the cafeteria staff are plausibly expressing, at least implicitly, that the prisoners have reason to drink. This can be seen though drawing a comparison to *Salience Nudge.* I suspect that we take *Salience Nudge* to implicitly express that the prisoners have a reason to choose the healthy foods because the placement of food at eye level is motivated in part by judgment that the prisoners have this reason. An analogous point holds also in respect of *Thirst Drug*. Here, the intervention is motivated by the thought that the prisoners have reason to drink water.

So it seems that *Thirst Drug* expresses that the prisoners have reason to drink water. Is this reason communicated, implicitly, to the prisoners? That is, do the prisoners *recognise* it? It is quite possible to specify the case in such a way that it is. We could, for instance, stipulate that *Thirst Drug* is subject to a kind of rational filtering analogous to that I described for nudges above: perhaps *Thirst Drug* works precisely by stimulating the prisoners to think carefully about whether they ought to drink water. Alternatively, we could posit a biological reason-tracking mechanism. We might, for example, simply stipulate that the thirst-inducing drug is much more effective at inducing prisoners to drink water the more dehydrated those prisoners are. The prisoners' responses to this intervention thus closely track their level of dehydration. And their dehydration also closely tracks—we may assume—their reasons to drink. Either way, the prisoners in *Thirst Drug* will plausibly count as being sufficiently sensitive to, and thus implicitly recognising, the reason expressed by the cafeteria staff. *Thirst Drug* will thus communicate—and so give—a reason. Yet it remains doubtful that *Thirst Drug* treats its targets as rational.

10. CONCLUSION AND IMPLICATIONS

Schmidt argues that some nudges neither subvert rationality nor harness irrationality. Levy argues that nudges typically give reasons. I have extrapolated these defences of nudges to some nonconsensual neurointerventions: Schmidt's argument implies that *Thirst Drug* neither subverts rationality nor harnesses

irrationality, and Levy's argument, on its most plausible interpretation, implies that *Thirst Drug* gives reasons if, for example, its effectiveness closely tracks the prisoners' levels of dehydration.

What further conclusions can we draw? It does not straightforwardly follow that some neurointerventions treat their targets as rational, for there may be other ways—ways not explored by Schmidt and Levy—in which an intervention can fail to treat its targets as rational. Schmidt and Levy take themselves to have undermined the most plausible bases for the objection that nudges fail to treat nudgees as rational, but they may have missed others. Moreover, there may be rationality-based objections to nonconsensual neurointerventions that do not plausibly apply to nudging and so are, understandably, not considered by Schmidt and Levy.

However, that an influence operates by giving reasons is standardly taken to be sufficient to establish that it, at least presumptively, treats the influencee as rational (Tsai 2014, 78-9).[32] Indeed, in the philosophical literature, giving reasons is frequently discussed primarily as a point of contrast for forms of influence, such as manipulation, that paradigmatically fail to treat as rational.[33] Thus, if *Thirst Drug* operates by giving reasons, that at least constitutes a defeasible case for the view that it treats its targets as rational. And if it also neither subverts rationality nor harnesses irrationality, that will only make the case harder to defeat.

This creates two challenges. The first is a challenge for opponents of neurointerventions. The challenge is to answer this defeasible case for the view that nonconsensual neurointerventions sometimes treat their targets as rational. This could be done by undermining the case—by showing that Schmidt, Levy or I have gone wrong somewhere, and that nonconsensual neurointerventions do invariably subvert rationality, harness irrationality or fail to give reasons. Or it could done by defeating the case—by showing that, despite neither subverting rationality, nor harnessing irrationality, nor failing to give reasons, nonconsensual neurointerventions do, in some other way, fail to treat as rational.

The second challenge is a challenge for Schmidt and Levy, and for those who endorse their defences. If my argument succeeds, these defences together generate a defeasible case for the view that some nonconsensual neurointerventions treat their targets as rational. But many will, I suspect, find this view counter-intuitive. Thus, unless the case can be defeated, a *reductio* threatens. Two routes are

---

[32] Tsai argues that this presumption can be overridden He argues that, though reason-giving influences are typically morally innocuous, they are not *invariably* so—some are in fact paternalistic since they 'occlude an opportunity for someone to canvass and weigh reasons for herself on her own terms' (p. 93) and are motivated by, and communicate, a negative judgment regarding the target's agency.

[33] See, for example, Berofsky (1983: 311); Shiffrin (2000: 213); Quong (2011: 81); Blumenthal-Barby (2012: 351).

available to blocking this *reductio*. First, nudge defenders could furnish an explanation of why nonconsensual neurointerventions, but not those nudges that they defend, subvert rationality, harness irrationality, fail to give reasons, or otherwise fail to treat their targets as rational. Second, they could concede that nonconsensual neurointerventions *do* sometimes treat their targets as rational, but deny that this implication is unacceptable.

COMPETING INTERESTS

The author has no competing interests to declare.

ACKNOWLEDGMENTS

I thank, for their funding, the Uehiro Foundation on Ethics and Education, and the European Research Council [grant number 819757]. For their comments on earlier versions of this manuscript, or discussion of the ideas contained therein, I thank Gabriel De Marco, Maximilian Kiener, Andreas Schmidt, Neil Levy, an audience at the Oxford-Zurich Work-in-Progress Meetings on the Ethics of Behavioural Prediction and Influence, and two anonymous reviewers for *Ethical Theory and Moral Practice*. For her research assistance, I think Tess Johnson.

REFERENCES

Anderson, Joel 2010. Nudge: Improving Decisions about Health, Wealth, and Happiness, Richard H. Thaler and Cass R. Sunstein. Yale University Press, 2008. x + 293 pages. [Paperback edition, Penguin, 2009, 320 pages.], *Economics and Philosophy* 26/3: 369–76.

Barton, Adrien, and Till Grüne-Yanoff 2015. From Libertarian Paternalism to Nudging—and Beyond, *Review of Philosophy and Psychology* 6/3: 341–59.

Bennett, Christopher 2018. Intrusive Intervention and Opacity Respect, in *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*, eds. David Birks and Thomas Douglas, Oxford: Oxford University Press: 255-73.

Berofsky, Bernard 1983. Autonomy, in *How Many Questions? Essays in Honor of Sidney Morgenbesser*, eds. Leigh S. Cauman, Isaac Levi, Charles Parsons, and Robert Schwartz, Indianapolis: Hackett: 301–20.

Birks, David, and Thomas Douglas (eds) 2018. *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*, Oxford: 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 Aanand D. Naik, 2015. In Defense of Nudge–Autonomy Compatibility, *American Journal of Bioethics* 15/10: 45–7.

Bovens, Luc 2009. The Ethics of Nudge, in *Preference Change: Approaches from Philosophy, Economics and Psychology,* eds. T. Grüne-Yanoff and S.O. Hansson, Berlin and New York: Springer: 207–19.

Bruns, Hendrik, Elena Kantorowicz-Reznichenko, Katharina Klement, Marijane Luistro Jonsson, and Bilel Rahali 2018. Can Nudges be Transparent and Yet Effective?, *Journal of Economic Psychology* 65: 41–59.

Bublitz, Jan Christoph 2018. The Soul is the Prison of the Body: Mandatory Moral Enhancement, Punishment, and Rights Against Neurorehabilitation, in *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice,* eds. David Birks and Thomas Douglas, Oxford: Oxford University Press: 289-320.

Forsberg, Lisa 2021. Anti-Libidinal Interventions and Human Rights, *Human Rights Law Review* 21/2: 384–408.
Cohen, Shlomo 2013. Nudging and Informed Consent, *American Journal of Bioethics* 13/6: 3–11.

Conly, Sarah 2013. *Against Autonomy: Justifying Coercive Paternalism*, Cambridge: Cambridge University Press.

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

Gorin, Moti 2014a. Towards a Theory of Interpersonal Manipulation, in *Manipulation: Theory and Practice*, eds. C. Coons and M. Weber, Oxford: Oxford University Press.

Gorin, Moti 2014b. Do Manipulators Always Threaten Rationality?, *American Philosophical Quarterly* 51/1: 51–61.

Grill, Kalle 2014. Expanding the Nudge: Designing Choice Contexts and Choice Contents, *Rationality, Markets and Morals* 5/90: 139–62.

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

Kroese, Floor M., David R. Marchiori and Denise T. D. de Ridder 2016. Nudging Healthy Food Choices: A Field Experiment at the Train Station, *Journal of Public Health* 38/2: e133–7.

Levy, Neil 2017. Nudges in a Post-Truth World, *Journal of Medical Ethics* 43: 495-500.

Levy, Neil 2018. Nudges to Reason: Not Guilty, *Journal of Medical Ethics* 44/10: 723.

Levy, Neil 2019. Nudge, Nudge, Wink, Wink: Nudging is Giving Reasons, *Ergo* 6/10, doi:10.3998/ergo.12405314.0006.010.

Madrian BC, Shea DF (2001) The Power of Suggestion: Inertia in 401 (k) Participation and Savings Behavior. The Quarterly Journal of Economics 116:1149–1187

McKenzie, Craig R. M., Michael J. Liersch, and Stacey R. Finkelstein 2006. Recommendations Implicit in Policy Defaults, *Psychological Science* 17/5: 414-20.

Pugh, Jonathan, and Thomas Douglas 2017. Neuro-Interventions as Criminal Rehabilitation: An Ethical Review, in *The Routledge Handbook of Criminal Justice Ethics,* eds. J. D. Jacobs and J. Jackson, London: Routledge.

Quong, Jonathan 2011. *Liberalism without Perfection*, Oxford: Oxford University Press.

Rozeboom, Grant J. 2020. Nudging for Rationality and Self-Governance, *Ethics* 131/1: 107–21.

Ryberg, Jesper 2018. Neuroscientific Treatment of Criminals and Penal Theory, in *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal*

Justice, eds. David Birks and Thomas Douglas, Oxford: Oxford University Press: 177-95.

Ryberg, Jesper 2019. Chapter 4: Neurointerventions as Punishment, in *Neurointerventions, Crime, and Punishment: Ethical Considerations*, Oxford and New York: Oxford University Press.

Schmidt, Andreas T. 2019. Getting Real on Rationality—Behavioral Science, Nudging, and Public Policy, *Ethics* 129/4: 511–43.

Shaw, Elizabeth 2018. Counterproductive Criminal Rehabilitation: Dealing with the Double-Edged Sword of Moral Bioenhancement via Cognitive Enhancement, *International Journal of Law and Psychiatry* 65/101378, doi: 10.1016/j.ijlp.2018.07.006.

Sher, Shlomi, and Craig R. M. McKenzie 2006. Information Leakage from Logically Equivalent Frames, *Cognition* 101: 467-94.

Shiffrin, Seana Valentine 2000. Paternalism, Unconscionability Doctrine, and Accommodation, *Philosophy & Public Affairs* 29/3: 205–50.

Steffel, Mary, Elanor F. Williams and Ruth Pogacar 2016. Ethically Deployed Defaults: Transparency and Consumer Protection through Disclosure and Preference Articulation, *Journal of Marketing Research* 53/5: 865–80.

Thaler, Richard H., and Cass R. Sunstein 2003. Libertarian Paternalism, *American Economic Review* 93/2: 175-9.

Thaler, Richard H. and Cass R. Sunstein 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New Haven, CT: Yale University Press.

Tsai, George. 2014. Rational Persuasion as Paternalism, *Philosophy & Public Affairs* 42/1: 78–112.

Wilkinson, T.M. 2013. Nudging and Manipulation, *Political Studies* 61/2: 341–55.

Wilkinson, T.M. 2017. Counter-Manipulation and Health Promotion, *Public Health Ethics* 10/3: 257–66.