Neuroethics (2024) 17:6
https://doi.org/10.1007/s12152-023-09540-3

ORIGINAL PAPER

What (if anything) morally separates environmental from neurochemical behavioral interventions?

Viktor Ivanković

Received: 10 March 2023 / Accepted: 2 November 2023 / Published online: 22 December 2023

Abstract Drawing from the literatures on the ethics of nudging and moral bioenhancement, I elaborate several pairs of cases in which one intervention is classified as an environmental behavioral intervention (EBI) and the other as a neurochemical behavioral intervention (NBI) in order to morally compare them. The intuition held by most is that NBIs are by far the more morally troubling kind of influence. However, if this intuition cannot be vindicated, we should at least entertain the *Similarity Thesis*, according to which EBIs and NBIs share relevant moral features to the extent that moral conclusions about one are implied about the other in the described pairs of cases. I test this thesis by putting forward a number of possible moral grounds for setting EBIs and NBIs apart, including three of the most promising ones – physical invasiveness, disclosure and avoidance, and inevitability. I conclude that although these promising grounds might not bear the full burden of vindicating the intuition against *Similarity* by themselves, clustering them together can establish discernible moral separation.

Keywords Nudges · Neurointerventions · Bioenhancement · Arational influence · Bypassing reasoning · Behavioral modification

V. Ivanković (✉)
Institute of Philosophy, Zagreb, Ulica Grada Vukovara 54, 10000 Zagreb, Croatia
e-mail: viktor.ivankovic@ifzg.hr

The Similarity Thesis

Consider the following pairs of cases:

Diners

> To gain an edge over their competitors from across the street, employees of Diner A change the color of its interior walls from a hunger-stifling blue to a hunger-inducing red [1, 2]. Their competitors (Diner B) counter this move by offering their customers complimentary drinks with a very small dose of Ghrelin, an appetite-increasing ligand [3, 4].

Cafeterias

> Two university cafeterias aim to encourage healthy food choices among their students. The staff of Cafeteria A utilize findings [e.g., 5] showing that food placed at eye-level is more salient and more likely to get picked; the staff place vegetables and fruits at eye-level. The staff of Cafeteria B spray the air of their cafeteria with a newly discovered substance that, once inhaled, creates a mild desire for flavors present in vegetables and fruits.[1]

[1] The Cafeteria B example is adopted from Douglas [6].

Examination rooms

> Motivated by the well-being of her patients, and convinced that she is offering them sound advice, Physician A consults research [7] on the kind of attire that is most likely to encourage trust and a cooperative attitude in patients, in order to procure their consent to her recommendations; as a result, she wears a white coat and tailored trousers in the examination room. Her colleague, Physician B, on the other hand, sprays her examination room with odorless oxytocin, a hormone known to increase trust and cooperation among people within social groups.[2]

City policies

> Cities A and B are faced with recent spikes in violent crime and suicide cases. Consulting recent research that exposure to green spaces decreases suicide mortality [9], City A significantly increases the combined acreage of green spaces (grass, parks), and in such a way that citizens can hardly avoid them while traversing the city. City B, on the other hand, decides to confront the negative trend by adding small quantities of lithium, which has been found to reduce suicidal and aggressive behavior, into the water supply [10, 11].

Our conundrum is the following: Are there any relevant moral differences between influences utilized by A-agents as opposed to those by B-agents? What, if anything, can explain our intuition that the latter influences are by far more morally troubling?

Let's add a few general stipulations to these cases. First, assume that all stated influences on behavior are not hazardous to health, or do not radically affect preferences and behavior that are not targeted by the influences. For instance, there would be no danger that inhaling oxytocin or imbibing lithium would not only increase one's trust or attenuate aggression respectively, but would also, say, ruin one's passion for the music of Ludwig van Beethoven. Second, suppose, for the time being, that targeted individuals in each case are unaware of the presence of the particular behavior-affecting interference, and indeed of each general kind of influence. For example, they are unaware that restaurant colors and drink additives are sometimes used to enhance appetite, and that they are indeed being used on that particular occasion.

Many influences of the former kind – those employed by the A-agents – have in recent years been assessed under the 'nudge' label [12]. 'Nudging' is the practice of designing environments in which choices are made in order to produce predictable effects on behavior, but without changing the content of choice options. It rests on findings from cognitive and behavioral sciences, particularly the heuristics and biases literature [e.g., 13]. Nudges were originally presented to policy-makers as a novel, easily resistible intervention for improving the welfare of targeted individuals [12, 14], but have over time also been considered for the facilitation of other goals, such as charitable giving [15], organ donation [16], and the discharging of enforceable duties, such as tax compliance [17]. Thus, nudges are now often conceived as boosting compliance with either prudential or moral reasons. Some may question whether all influences utilized by A-agents above are nudges proper; some are not mild or do not target a particular decision (the green spaces in *City policies* are intended to produce a more general behavioral change), while others do not boost compliance with either moral or prudential reasons (the red walls in *Diners* are intended instead to benefit the influencer). However, not only do these influences resemble the paradigmatic nudges in the literature (changes in the physical environment, e.g., in traffic), but they also appeal to findings according to which behavior is explained without reference to the slow and deliberate System-2 processes of reasoning. Moreover, these influences will likely be objected to on virtually identical grounds as nudges, for example, that they are manipulative [18] or fail to treat people as rational agents [19]. Without committing to the view that each of these influences is a nudge in the strictest sense, I draw on the nudge literature to gather at least some moral lessons about these and similar heuristic triggers, which I will here refer to as 'environmental behavioral interventions' (EBIs).

Some influences of the kind exerted by B-agents in my examples have in turn been morally explored in recent years under the 'moral bioenhancement' label. This debate has assessed whether neurochemical interventions upon brain states, like the administration

[2] See Gelfand [8] for a similar comparison.

of the trust-inducing hormone oxytocin [20, 21], or the fairness- and cooperation-promoting neurotransmitter serotonin [22], can ever permissibly be used to enhance deficient moral dispositions, such as our lacking propensities for fairness, altruism, or our faltering adherence to important social norms. The non-consensual use of neurointerventions has only rarely been entertained (e.g., in the context of criminal rehabilitation [23]) and has come under critical fire when advocated for the prevention of catastrophic harm to humanity (for advocates, see [24, 25]; for critics, see [26, 27]). As with the nudge literature and EBIs, the moral bioenhancement literature will be instructive for the influences I explore here under the 'neurochemical behavioral interventions' (NBIs) label.

Although they do so in different ways – the former by triggering heuristics and the latter by more directly influencing neurochemical processes – EBIs and NBIs in the pairs of cases above both alter psychological states and/or modulate emotions by bypassing human capacities for explicit rational deliberation. Thus, they raise several moral concerns, such as that they may reduce the responsiveness of one's motives to reasons for or against an action [28], or that they may expose agents to the domination of others [29, 30]. And yet, even if both EBIs and NBIs are deemed morally problematic, an intuition might persist that NBIs are *more* threatening. Should we hold onto this intuition? Can we vindicate it by identifying a morally relevant difference that it may be tracking?

A first stab at vindicating the intuition might be to point to a common notion that most EBIs are *mild* influences. Some have remarked that, as policy tools, nudges in particular may be no better than techno-fixes that "cannot solve complex policy problems" [31], and by themselves cannot tackle a burning issue like climate change [32]. Even their founders, Richard Thaler and Cass Sunstein, have stated that for such problems, "gentle nudges may appear ridiculously inadequate – a bit like an effort to capture a lion with a mousetrap" [12]. Conversely, moral bioenhancements have been advocated for their supposed capacity to tackle exactly these kinds of existential problems. In a rare reference to nudges in the moral bioenhancement literature, Ingmar Persson and Julian Savulescu have claimed that, due to their mildness, nudges "are not well suited to induce behavioural changes that should be radical and permanent" [24].[3] Consider also that nudges are defended on grounds that they are a kind of influence that preserves liberty [12], whereas one of the most oft-cited charges against moral bioenhancements is that they vitiate the freedom of autonomous agents [26].

However, I contend that an appeal to mildness should not suffice. Even if we imagine that EBIs from my introductory pairs of cases produced effects that were *just as strong*, or even *somewhat stronger* than those by their NBI counterparts, the intuition that NBIs are more morally troubling would likely persist. For instance, we would likely take autonomy to be more seriously threatened by Cafeteria B, where a drug is sprayed into the air, than by Cafeteria A, where healthy foods are placed at eye level. Similarly, most would take the administration of oxytocin spray in an examination room to be more threatening to autonomy than a physician wearing a white coat and tailored trousers, even if these influences were equivalent with respect to the magnitude of behavioral effects. I will thus proceed with a further stipulation – that the influences within the comparisons are roughly equal with respect to the magnitude of their effects, i.e., that one of the influences is not significantly more potent upon taking effect. For instance, imagine that, in *Diners*, Ghrelin can be dosed so that it produces a similar change in appetite as red walls, and that the enhanced hunger is equally difficult to suppress and resist in both cases. Hence, to vindicate the aforementioned intuition, we need to identify morally relevant features that would set EBIs apart from NBIs other than magnitude of effect. If we cannot, then we should at least entertain the following:

> *Similarity Thesis*: Holding fixed the magnitude of their behavioral effects and the absence of side-effects, EBIs and NBIs share relevant moral features to the extent that any moral conclusions we may draw about EBIs should be implied about NBIs and vice versa.

If *Similarity* holds, then any arguments made in favor or against one kind of influence in the pairs could, with equal success, also be made about the other

[3] See also Schaefer's claim that the greater efficiency of NBIs compared to EBIs is owed to their greater capacity to "isolate particular psychological functions" [33].

kind.[4] Defending EBIs, on the one hand, would commit one to defending 'similar' NBIs that intuitively seem more morally threatening. Objecting to NBIs, on the other hand, would commit one to applying that same criticism to 'similar' EBIs that intuitively seem much less morally threatening. In a way, this sets up a challenge for EBI proponents to look for relevant differences, and for NBI proponents to deny them, or to reinforce the supposed similarities. If, however, EBIs and NBIs are dissimilar, we are not necessarily left with the straightforward conclusion that NBIs are morally threatening and impermissible, whereas EBIs aren't; NBIs could simply be *more* threatening comparatively while both are impermissible, or both are permissible. But a failure of *Similarity* could show that, at least in cases where we would consider influencing people with both kinds of influences to be all-things-considered permissible, EBIs should be the first, more preferable option.[5]

The moral debate on *Similarity* is still in its infancy. Thomas Douglas [35] has argued that it's hard to find any property that explains why moral intuitions decisively favor EBIs over NBIs. In his most recent articles, he argued that arguments often stated in favor of EBIs – pertaining to mere substitution of influence [6] and the treatment of targeted agents as rational [36] – similarly apply to NBIs. Along similar lines, Jonathan Pugh suggested that any objection to NBIs must show how they threaten freedom in a way that EBIs don't [37]. Contrary to their claims, Jan Christoph Bublitz and Reinhard Merkel believe targeted individuals have more control over EBIs insofar as they are mediated by perceptual processes [4].

Here I attempt to extend the analysis of *Similarity* by looking at existing and further, as-yet-unconsidered grounds that may set EBIs and NBIs apart. To the question whether such relevant distinguishing features can be identified, I offer a very cautious 'yes'. In the next three sections, I test what I consider to be the three main candidates for establishing this moral distinction, pertaining to physical invasiveness, disclosure and avoidance, and inevitability. On these grounds, we can draw what I take to be the most convincing arguments for morally separating EBIs and NBIs. I then briefly elaborate some other candidates that at least merit a mention. The concluding section considers the cluster view, according to which the intuition that NBIs are significantly more threatening could only be sustained by considering more than a single moral consideration.

Physical Invasiveness

Intuitively the strongest candidate for setting apart EBIs and NBIs morally is that NBIs are physically invasive to their targets. In all examples introduced above, the chemical agents in the NBI category must enter the bodies of their targets, by being inhaled or imbibed, in order to introduce changes in brain chemistry. These are evocative of cases where one is drugged by another, without consent or awareness, to make it more likely that she will act in accordance with the latter's desires. EBIs, on the other hand, presumably only modify the environment in which options are represented, and hence do not violate the body. Thus, as Douglas points out, NBIs, and not EBIs, might be regarded as violating the right to or interest in freedom from bodily interference [35]. Taking bodily integrity seriously in these kinds of cases entails giving individuals a say in what substances enter their bodies. Since, by assumption, NBIs like lithium in *City policies* and the oxytocin spray in *Examination rooms* are added without being disclosed to those they target, a right to bodily integrity that is taken seriously would entail that individuals can raise reasonable claims against being subjected to such interventions.

[4] *Similarity* should be distinguished from Levy's parity principle [34], which states that the moral permissibility of these interventions should only be assessed in terms of their costs and benefits, and not their natures. On the *Similarity Thesis*, the natures of interventions are not ignored; the moral permissibility of both EBIs and NBIs should still be determined not only on the basis of outcomes, but also on the basis of the kinds of interventions that they are. *Similarity* only asserts that their natures are morally indistinguishable and warrant the same conclusions about their permissibility.

[5] Note that we might reach different conclusions about different pairs of EBIs and NBIs in the examples, either because of the goal that they promote or because the particular EBIs (and/or NBIs) are morally dissimilar between themselves. I will point out explicitly in what follows when a particular distinguishing feature is more significant for some pair of cases than for another.

I don't wish to deny that physical invasiveness of NBIs is to be listed among the morally relevant distinguishing features that undermine *Similarity*. Instead, I suggest that the capacity of physical invasiveness for this purpose might be more limited than is at first intuited. Despite its obviousness, it will hardly bear the full burden of vindicating the intuition against *Similarity* by itself. I offer three reasons to think of physical invasiveness as a relevant, but limited distinguishing feature.

First, Douglas convincingly points out that the more morally significant aspect of NBIs seems to be the mental interference that they involve [35]. If so, and if we maintain that EBIs and NBIs are equivalent in magnitude of effect, then EBIs arguably involve a 'similar' mental interference as NBIs. Of course, similarity in terms of mental effect doesn't mean we have to give up on the appeal to physical invasiveness; EBIs presumably *only* interfere with targets mentally, while NBIs interfere both mentally *and* physically. Nevertheless, the mental effect seems to be the more significant aspect of the interference. As Douglas points out, if we were to imagine a physically similar influence (say, a spray) that didn't affect mental life, but instead stopped the spread of viral infections, we wouldn't find it as objectionable as the cafeteria spray or the oxytocin spray [35]. Similarly, Bublitz and Merkel have offered examples much like the NBIs in *Diners* and *Examination rooms*, which they claim "are potentially illegitimate only in virtue of their mental effects", and would not even constitute a violation of bodily integrity [4]. To be sure, Bublitz and Merkel's denial of NBIs' violation of bodily integrity seems exaggerated here, but it does seem to support the intuition that the mental effect constitutes the more relevant aspect of NBIs' interference. Now, if the physical interference is indeed less significant, it might still be difficult to offer any kind of philosophical reasoning of just how the aspects contained in the same interference can be compared. Philosophical reasoning may only be able to offer attempts at substantiating the point about significance qualitatively, like in the fact that the upset of mental life is an intended effect of an NBI, whereas the upset of physical life is its foreseeable, but unintended side-effect.[6]

Second, while EBIs are not physically invasive in the standard sense, we might want to contemplate ways in which they might be considered physical or invasive in non-conventional senses. Let's consider the physical properties of EBIs. These interferences physically affect the neural states of their targets through visual, auditory, olfactory, and conceivably, tactile stimuli. In the nudge literature, examples of visual interferences seem like the most common, auditory less common, while olfactory are hardly ever mentioned. So why don't we conceive of EBIs as physical interferences as we do of NBIs? Part of the explanation is surely that, conventionally, we do not think of light particles *entering* the eye or auditory waves *entering* the ear as the same kind of physical interaction with the body as chemical agents entering the body through the air or the water supply. This, in turn, might be explained by the fact that, unlike chemical substances of the kinds described, light particles and auditory waves interact with our senses continuously, so tweaking them hardly seems to introduce a new kind of sensory interaction that would constitute a physical violation.[7] Yet, we may think of olfactory stimuli differently. Drawing on research by Morrison et al. [39] on the effects of aroma influences on behavior, Scott Gelfand imagines an influence exactly like the NBI in *Examination rooms*, except that instead of an oxytocin spray, a vanilla aroma is used to produce the same effect: namely, the encouragement of trust and a cooperative attitude in patients [8]. Despite the vanilla aroma 'merely substituting' smells in the environment, and despite being mediated through perception and/or other psychological mechanisms, it seemingly interferes with its targets physically the same way as the oxytocin spray. If so, it would be difficult to say why the oxytocin spray violates physical integrity, and the vanilla aroma doesn't. As for other EBIs, those mediated through vision and hearing, we would need an account of the kind of physical interaction amounting to a violation of physical integrity that explains why visual and auditory EBIs do not constitute such violations.[8] Before coming up with such

[6] However, the use of the intend/foresee distinction in characterizing NBIs' effects could also be contested here, e.g., by noting that the physical interference seems *constitutive* to the intended effect, rather than an unintended side-effect [38].

[7] See more about the mere substitution of influence in Section "Inevitability".

[8] Alternatively, one might argue that the vanilla aroma is an NBI, not an EBI, but that comes at the cost of the distinctions between EBIs and NBIs on the one hand, and between direct and indirect interferences on the other, becoming seriously disjointed. More about the latter distinction in the next section.

an account, we should remain at least mildly agnostic about which kinds of physical interaction constitute violations of physical integrity.

Additionally, the way in which some EBIs physically interact with individuals should be considered invasive, especially if they render their targets incapacitated in pursuing some course of action, or cause pain or discomfort. A particularly bright light or a foul odor may have this effect, and may even be used for purposes of torture. Such cases of EBIs would likely be considered at least as invasive as stated examples of NBIs. However, the invasiveness threshold would likely be crossed only by the cases of EBIs that incapacitate or severely discomfort their targets, whereas the threshold for the invasiveness of NBIs is seemingly much lower and crossed by virtually all NBIs, regardless of whether they leave their targets incapacitated. In short, then, there are cases of EBIs that interact with their targets in a physically invasive way, but these are much rarer than cases of NBIs.

Finally, even if some EBIs interfere with their targets physically the same way as NBIs, it's not entirely clear why that makes them objectionable; and we should know why if that very same feature adds to the objectionability of NBIs. Consider again Gelfand's example of the vanilla aroma. While administering the aroma covertly might constitute an objectionable mental interference, it's not entirely clear why physically inhaling a pleasant vanilla scent would add much, if anything to this violation. The interference seems just as objectionable as other EBIs, like the red walls in *Diners* or the green spaces in *City policies*, and is possibly much less objectionable than some other EBIs that are not physically invasive in the standard sense, like the use of subliminal messages that motivate pro-social attitudes. Arguably, then, the physical interference by some EBIs might not be relevant morally as it supposedly is in cases of NBIs.

Let's take stock. Physical invasiveness is the most intuitive candidate for undermining *Similarity*, but the strength of this feature is difficult to fully vindicate. The appeal not only 1) seems less significant than the mental interference that these influences cause, but 2) it remains uncertain what conditions need to obtain for different kinds of physical interactions to violate physical integrity and 3) it remains uncertain whether the appeal does the same moral work when an EBI physically interferes with targets in a seemingly similar way. While physical invasiveness will find its place among features that provide some moral separation between EBIs and NBIs, it's questionable whether it can fully vindicate the intuition against *Similarity* by itself.

Disclosure and Avoidance

I next consider Bublitz and Merkel's proposal that NBIs are more threatening than EBIs in virtue of targeted individuals having less control over them [4]. On their view, (most) EBIs are *indirect* interventions, in the sense that they are mediated through some aspect of perception and through psychological mechanisms (many of which EBIs trigger to produce some desired effect). Conversely, NBIs are *direct* interventions, in the sense that they do not operate via mediating psychological processes when altering brain states. This is not to suggest EBIs come anywhere close to being fully controllable, as some of them never rise to the level of their target's conscious awareness. Still, Bublitz and Merkel hold that mediating processes at least tentatively ensure targets more control over EBIs than over NBIs [4].[9]

Douglas resists this proposal on three grounds [35]. First, he writes that we could easily imagine NBIs to be visible to their targets (e.g., for spray particles to be visible in *Cafeterias*). Although this wouldn't make perception constitutive to the production of behavioral modification in the case of NBIs, Douglas notes, it would introduce perception as a mediating point, which may allow for individuals to monitor NBIs in the same way as it's possible for EBIs. Additionally, Douglas notes that making NBIs visible doesn't allow targets to monitor the *process* by which behavior is modified, but the same seems to apply to visible EBIs; in fact, the mental processes by which behavior is modified seem similarly opaque to

[9] Note that the distinction between direct and indirect interventions doesn't always fit neatly with the distinction between EBIs and NBIs. For instance, if subliminal effects were used to garner sympathy among targeted individuals for some pro-social cause (see [40]), then such influences would not involve any neurochemical effects on the brain, and would thus more likely be classified as EBIs, yet they would, in Bublitz and Merkel's words, only be using "peripheral routes of perception" [4] and would arguably not be mediated via perception as most other EBIs. Still, the two distinctions will fit each other neatly in most other cases.

targets for both EBIs and NBIs. Second, the thought that perceptual mediation allows for easier resistance of the influence, Douglas believes, also seems unwarranted. For instance, a person exposed to red walls in *Diners* may have to take extreme measures, such as keeping her eyes closed at all times, to block their effect. Finally, perceptual mediation, according to Douglas, hardly entails that EBIs won't take effect (e.g., that red walls won't induce hunger); their effects could conceivably be overridden, but the same is conceivable for NBIs as well, holding fixed the magnitude of effects.

But I now want to consider two different ways in which the indirectness of EBIs could be morally significant.

Consider first that triggering heuristics or some uncategorized psychological mechanism is *constitutive* for EBIs in producing their effects. If targeted individuals could prevent some such heuristics and mechanisms from being triggered at the level of conscious awareness (the state in which they, according to Bublitz and Merkel, have the most control over influences), then they could prevent some EBIs from ever taking effect on their mental states. This seems to be driving the intuition that nudges are more effective as long as they are non-transparent. For instance, Luc Bovens says that nudges, in some of their forms, "work better in the dark", and that making particular instances of nudging transparent would dull their effect because they owe their effectiveness to their covertness [41]. Similarly, Till Grüne-Yanoff states that because nudges rely on psychological quirks, they "will be more effective if they are not transparent to the individuals subjected to them" [18]; he also suggests people "will no longer find the drastic slogans and images [on cigarette packs] shocking" once they understand them, and that their "effectiveness […] requires their being not fully transparent" [18].

If transparency may change the way preferences are formed at the outset of exposure to EBIs, then the effect of prior disclosure about interventions on preference formation could be an autonomy-relevant feature that sets EBIs apart from NBIs. Consider *Cafeterias* and *Examination rooms*. It seems intuitive that an individual who knows about the workings of cafeteria nudge or physicians wearing white coats could sidestep their effect on behavior in virtue of this knowledge. It's questionable in such a case whether her heuristics and other psychological mechanisms are ever triggered. But if the same individual knew about the effects of cafeteria spray or examination room spray and found herself in the earlier described contexts, avoiding their influence *entirely*, as in the case of EBIs, would plausibly only come at the cost of extreme measures of the kind Douglas mentions, like holding one's breath or wearing a gas mask. Seemingly, their effect on preference formation cannot merely be sidestepped, but can be resisted only after it has influenced mental processes. Whereas prior disclosure of EBIs allows agents to prevent heuristics from being triggered, the direct character of NBIs leaves no room for such prevention. In other words, disclosures of EBIs afford agents a kind of ability to anticipate and resist that is not afforded by disclosures of NBIs. That at least is my tentative claim.[10]

Note that nudge theorists are cautious about the affordances of prior disclosure. Nudges are only presumed to work 'better' in the dark, or to be 'more effective' if covert.[11] Indeed, many EBIs seem difficult to disarm even when made transparent. For instance, red walls in *Diners* and green spaces in *City policies* seem like the more difficult influences to sidestep than EBIs in *Cafeterias* and *Examination rooms*. This is because we are permanently exposed to these influences – they keep "coming back at us" – and we have to mentally process their effects

[10] In a hypothetical example, Coons and Weber say that if we were forcibly injected with a love potion that has its standard imaginary effects, and we were told about this beforehand, disclosure would only make things worse for our agency as we would witness ourselves succumbing to its effects [42]. Intuitively at least, the love potion seems more analogous to NBIs than to EBIs. This is not to suggest that NBIs hijack agency in the same way as the love potion. However, agents might similarly witness their mental states being affected before coming to resist the NBI.

[11] One piece of empirical evidence seems to suggest that nudges need not lose their potency even if made transparent; namely, informing people that their choices regarding advanced directives are affected by defaults does not deter them from accepting the direction of the nudge [43]. Some might suggest that this means nudges still work even when disclosed. Two points should be made on this matter. First, it isn't clear that, following disclosure, the default in the study still works as a heuristic trigger, since those exposed to the default may have simply become aware of the reason behind it and have consciously endorsed it. In other words, it isn't clear whether behavior is still changed as a consequence of the behavioral technique being employed [44]. Second, it's not clear that what the study shows can be extended to other cases of triggering heuristics.

repeatedly. But there is at least some evidence that we may be able to nip some EBIs in the bud. Drawing on the empirical research of Miller and Fagley [45] and Sieck and Yates [46], Gregory Mitchell has argued that we can overcome the influence of EBIs to which we are exposed (e.g., "simply asking people to give reasons for their choices can reduce the influence of gain/loss effects") [47]. Or consider the research by Almashat et al. [48], who show that framing effects in the medical context can be eliminated by having patients fill out a simple questionnaire. Conceivably, other EBIs could be prevented by prior disclosure or by engaging reflective deliberation in some other way.[12]

The ability to sidestep effects as a result of disclosure might still be regarded by some as insufficient to sustain the original intuition that EBIs and NBIs are dissimilar, seeing that it might lack import in practical cases. Recall that in my stipulations, and often in real life cases, EBIs are non-transparent, so we can expect them to influence our preference structure in exactly the same way as NBIs. Absent a condition of transparency, it would seem, the fact that EBIs do not influence preferences directly, as NBIs do, makes no difference to which preferences targeted individuals end up having. And after all, if transparency were to block their effect, would there still be a point to utilizing EBIs?

I offer an example that solves both worries. Imagine a university informing students that it will administer health-promoting techniques, but will not disclose openly, through some public channel, what these techniques are and how they work. Instead, it will provide booklets at designated places at the university for interested individuals, specifying how the techniques work. Imagine now that the main technique is cafeteria nudge. This kind of setup affords those who consent to the promotion of health the opportunity to accept the nudge by staying oblivious about the details and cash in on the effects, while allowing dissenters to reject the influence by becoming familiar with the details of the nudge. Conversely, if the technique were cafeteria spray, the dissenters would not be able to fully avoid influences that pull them in the health-promoting direction.[13] Thus, the disclosure of EBIs grants dissenters more control over the intervention than the disclosure of NBIs.

A second way in which the supposed indirect character of EBIs may be morally relevant for avoiding influence concerns the social level. Consider Bovens' conception of watchfulness in a society utilizing nudges:

> "A watchful person would be able to identify the intention of the [nudge] and she could blow the whistle if she judges that the government is overstepping its mandate […] we stipulate that every *Nudge* should be such that it is in principle possible for everyone who is watchful to unmask the manipulation." [41]

What makes Bovens' watchfulness possible in the case of EBIs is mediation of some aspect of the EBI through perception. NBIs, on the other hand, which are not mediated through perception, have to be made visible, as in the case of visible particles of cafeteria spray, or some apparatus tracking their presence has to be made available for watchfulness to be possible. Otherwise, we cannot hope to contest their presence in a democratic setting. And even if such resources for monitoring NBIs were made available, in order to ensure watchfulness and democratic contestation, the monitoring of EBIs would likely be much less costly and more reliable, since it only requires the ability of some individuals to identify EBIs through perception and communicate their findings to the citizenry. Tracking a single NBI (let alone multiple ones) requires technological solutions that are not made readily and cheaply available (which will likely continue to be the case in the foreseeable future).[14] Thus, from my current point of view, EBIs seem like the more easily monitored, and thereby the more controllable kind of influence at the social level.

Once again, let's take stock. I have suggested two ways in which the indirectness of EBIs may be

[12] We should also be cautious about my assumptions regarding NBIs. Conceivably, their influence could be so mild that we hardly notice ourselves being affected before resisting them. But I want to retain the understanding that, as direct interventions, NBIs will produce some change in mental states before it can be resisted, slight as it may be.

[13] A similar strategy for the utilization of nudges is advocated in Ivanković and Engelen [44].

[14] Of course, this distinguishing feature may be more applicable to our current technological context. New technological solutions may narrow the gap between EBIs and NBIs in terms of the costs and reliability of tracking particular influences.

a morally relevant feature for undermining *Similarity*: 1) it seems plausible that disclosure will afford those exposed to at least some EBIs the opportunity to block their effects, by preventing heuristics from being triggered; the disclosure of NBIs does not leave room for such prevention, as individuals can resist their influence only after it has affected mental states; 2) Bovens' watchfulness, the ability to blow the whistle on some influences being used in order to contest their presence in a democratic setting, is achieved only in the case of EBIs because they are mediated through perception; to track and disclose NBIs in the same way would require making the interventions visible or providing a device that could track them.

Inevitability

The inevitability argument has featured prominently in accounts favoring the use of nudges, so much so that Kalle Grill has called it "Thaler and Sunstein's most important argument for nudging" [49]. It has been presumed to ward off most objections raised against nudging, including, for instance, that it treats its targets disrespectfully [50] or diminishes their control over their deliberation [51], among others. These objections don't get off the ground, the inevitability argument goes, because every choice environment is bound to *inevitably* influence those exposed to it in *some* way [12]. Choice architects are faced with inevitably setting up choice arrangements that will steer the individuals exposed in this or that direction; their nudges don't introduce a new kind of influence. In *Cafeterias*, it's hardly controversial that if the manager doesn't steer the students towards healthy food by making it visually salient, they will be influenced by a different food arrangement – one that is perhaps set at random, or one from which the manager aims to profit.

On Thaler and Sunstein's account, the inevitability argument has a welfarist flavor: if choice environments are bound to steer individuals' choices in *some* way, then it's best that they steer them towards decisions that maximize their welfare [12].[15] As I noted early in the paper, many advocates have since rejected the welfarist route, and there is no reason to normatively restrict the argument in such a way as we apply it more broadly to EBIs. In other words, the inevitability argument shouldn't limit the various normative alternatives that an influence might promote. Rather, it only places EBIs beyond the criticism that they're impermissible altogether. To illustrate this, consider the painted walls in *Diners*. The inevitability argument cannot insulate the A-agent from criticism that she shouldn't have painted the walls red, in order to boost sales – she could've indeed painted the walls in some other color and thus pursued a different moral direction. But the inevitability argument does insulate her from the criticism that she shouldn't have painted the walls to produce a behavioral effect altogether, since any color is bound to have *some* effect.

Douglas [6] claims that if the inevitability argument, or as he calls it, 'the Mere Substitution Defence', can be successfully invoked to defend the use of EBIs, then we should be able to invoke this argument to that same end for NBIs. NBIs merely substitute influences that already apply to choice; what is substituted are the chemical features of our brains and of the environment, affecting our motivations to make some choice rather than another. Douglas calls this the *motivational context* of choice [6]. Consider Douglas's own case of the cafeteria spray in *Cafeterias*. If individuals are not sprayed to opt for vegetables, then they will have neurochemically influenced motivations to pick some other food. In *Examination rooms*, if patients don't inhale the oxytocin sprayed in order to encourage their cooperativeness, they may, by sheer chemical accident, be influenced to act uncooperatively. The NBIs "simply replace one set of chemical influences with another" [6].

To be sure, the inevitability argument, either for EBIs or for NBIs, is by no means foolproof. Both proponents [e.g., 57, 58] and opponents [e.g., 18] of nudges have argued that there is an autonomy-relevant difference between accidental influences on choice and intentional interventions aiming at a particular behavioral effect.[16] Douglas's 'Mere

[15] Similar appeals to inevitability can be found in Sunstein [52, 53], Cohen [54], Brooks [55], and Engelen et al. [56].

[16] However, a more sophisticated, and, in my mind, quite convincing version of the inevitability argument, which states that there is little, if any moral difference between actively changing environments and allowing them to take effect when all likely outcomes are reliably predictable to the choice architect, is offered by Blumenthal-Barby [59, emphasis in original]: "once behavioral science helps us gain insight into how choice is affected, intentionality is forced, in a sense. It becomes increasingly difficult for us to maintain that we did not know how various factors in the choice architecture would impact

Substitution Defence' isn't meant as a full defense of either EBIs or NBIs by means of the inevitability argument; instead, it's meant to show that it's hard to see why it wouldn't be applicable to both, or neither. In other words, the claim is that, pertaining to the success of the inevitability argument, EBIs and NBIs stand or fall together. It should be the aim, then, of any opponent of *Similarity* to challenge the analogy between EBIs and NBIs with regard to the inevitability argument, i.e., to show why the argument may succeed for EBIs but not for NBIs.

The most promising strategy that could be pursued to this end, I believe, is to say that for EBIs that I mention here, there is an additional sense in which their performers may appeal to inevitability. I call this the constitutiveness argument. These substitutions of accidental influences are in a way *constitutive* to some activity being performed; the A-agent cannot avoid making choices that include them. Consider *Cafeterias*. When Cafeteria A arranges food for its customers to select, it's performing an act that seems constitutive to running a cafeteria – we can hardly imagine a cafeteria being run without the stage of arranging food in its operations. With knowledge about behavioral findings influencing choice, employees of Cafeteria A cannot avoid setting up some arrangement of food items that will predictably influence their customers. It's still possible for employees of Cafeteria A to wrong their customers, e.g., by setting up the cafeteria in a way that harms the health of choosers when they could've set it up otherwise, but it remains the case that they have to arrange it in *some* way. For Cafeteria B, the choice of administering the spray doesn't seem so constitutive. It's hardly the case that changing the precise neurochemical contents of the air is an unavoidable part of running a cafeteria. Similarly for *Examination rooms*, the decision whether to put on, say, casual clothing, scrubs, or a white coat, must be made daily by Physician A before entering the examination room and interacting with her patients – it cannot be avoided [see also 54]. Physician B, on the other hand, can avoid making decisions pertaining to the neurochemical contents of the air inside the examination room. The same argument, I believe, can be made for A-agents and B-agents in *Diners* and *City policies*, although in these cases, the inevitable decision is faced by A-agents less often, or only once. Hence, EBIs often seem like parts of inevitable decisions that must be made by their performers, unlike NBIs.

While this argument is very convincing, the difference might not hold in all cases. Some NBIs could be conceived so that they are constitutive to some activity being performed. Consider the decision between different ventilation systems for a building that will affect levels of oxygen, and thus levels of focus for the building's inhabitants. Since such a decision determines neurochemical contents that will predictably affect behavior, it wouldn't be too much of a stretch to characterize it as an NBI. If so, there are some NBIs to which the constitutiveness argument could apply. But notice that, at least for the choice of ventilation systems, our initial intuition that NBIs are more morally threatening seems to disappear – there seems to be nothing objectionable about installing a ventilation system in itself.[17] Hence, the constitutiveness argument provides opponents of *Similarity* with at least two strong points: 1) EBIs are often a constitutive part of inevitable decisions, unlike most NBIs, and 2) when NBIs are constitutive in such a way, they don't seem objectionable.

Noteworthy Mentions

So far, I have presented three of what I take to be the most promising features that could morally separate EBIs from NBIs and undermine *Similarity*. In this section, I turn to some of the less convincing arguments for undermining *Similarity*, but which may nevertheless warrant further consideration in future research.

Footnote 16 (continued)
[…] choice. […] Given that we then have to make a decision about *how* to set things up, we are forced to engage in nudging or shaping choice one way or the other".

[17] I thank Tom Douglas for putting this example forward in correspondence. Note that, as in the case of Cafeteria A, the decision-maker would be acting blameworthy if her choice of ventilation system was predictably harmful to the inhabitants, e.g., by making them docile. I thank an anonymous reviewer for raising this concern.

To start off, consider that NBIs mostly lack referents in used, or even prototypical techniques, insofar as they are linked to the techniques from the debate on moral bioenhancement, which tends to be mostly hypothetical. The early stage of development of these technologies is sometimes found to be at odds with the existential urgency that is thought to support the proposal for their use [60]. Others have more broadly criticized the project for being too speculative and based on question-begging assumptions [61].[18] On the other hand, EBIs have more realistic referents in techniques that are already utilized, namely nudges. If that's the case, shouldn't the greater moral threat reside in EBIs, which are readily available to policymakers today? Interestingly, if such an appeal to the realism of EBIs were to hold, it would represent an interesting anomaly: it would undermine *Similarity*, but in the "wrong direction" – it wouldn't be vindicating the intuition that NBIs are more threatening than EBIs, but would be proving the exact opposite. The appeal to realism certainly seems relevant for where we direct focus in our current context of various influences, but a thorough moral assessment of the two kinds side-by-side seems to stretch beyond that context, making it sensical for us to put the appeal aside. Also, how realistic particular interventions are may have to be determined on a case-by-case basis. For instance, the NBI in *City policies* may be more realistic than the one in *Cafeterias*. NBIs in *Diners* and *Examination rooms* may become more realistic in time.

Second, some might appeal to Neil Levy's argument that nudges do not bypass reasoning (so neither would EBIs by extension). According to Levy, nudges take effect "by giving us reasons": "The mechanisms that respond to nudges are reasoning mechanisms, and in most cases, at least, nudges no more bypass reasoning than do philosophical arguments" [63]. Because nudges tend to make certain options more salient to us, the argument goes, they seem to operate as implicit recommendations [63]. Since mechanisms that respond to NBIs are *not* reasoning mechanisms, Levy's argument could be invoked by opponents of *Similarity* as offering an important distinguishing feature. But Douglas [36] challenges this view by stating that if 'nudging giving reasons' merely consists in a cafeteria nudge implicitly expressing the proposition 'you have reasons to choose healthy foods', and in the salience of those foods sufficiently steering nudgees towards recognizing reasons to choose healthy foods, then some neurochemical interventions could also be said to give reasons. The objection contends that the spray in *Cafeterias* would similarly express, implicitly, that the customers have reasons to choose vegetables and fruits, and it could sufficiently steer them towards recognizing these reasons. Douglas thus faces Levy's view with an important challenge that ought to be addressed by those who want to suggest that Levy's notion of giving reasons in the case of nudges doesn't extend into neurochemical territory and can be used to set EBIs and NBIs apart. To me, at least, it's not obvious how one might go about responding to this challenge.

For the third and final noteworthy mention, consider that opponents of *Similarity* could appeal to David DeGrazia's view on alienating influences, according to which individuals are autonomous if they have preferences that they identify with, and they did not come to identify with these preferences as a result of influences that they consider alienating [64]. It's more likely, the appeal might go, that given our public culture and the intuition against *Similarity* that we have set out with, individuals will be more likely to approve EBIs as non-alienating and reject NBIs as alienating. If so, then *Similarity* is to be simply rejected in virtue of NBIs being considered an alienating kind of influence by those exposed to them; we do not need to seek out moral differences beyond this likely public opinion of NBIs. But, of course, it would have to be determined empirically whether NBIs would indeed be rejected as an alienating kind of influence. And even if they were, the particular instances of EBIs described here could also be rejected as alienating. Whether influences would be rendered alienating would likely depend on how familiar subjects are with them, and there is little reason to expect familiarity to neatly separate the two categories. Furthermore, the proponent of this possible difference would also need to offer a defense of DeGrazia's subjectivist view against the criticism of a more objectivist stance that some influences are alienating regardless of whether those exposed to them

[18] More specifically, for a criticism about the impracticability of SSRIs as a means of moral bioenhancement, see Wiseman [62].

believe so, for instance, in accordance with some of the distinguishing features that I have offered here.

Concluding Remarks

The three features that I elaborated in this paper as the most promising for morally separating EBIs and NBIs pertained to physical invasiveness, disclosure and avoidance, and inevitability. Some might object that though these are valid grounds for moral separation, they are not particularly powerful ones. If disclosure is absent, then there is no practical import to the insight about avoidance. The inevitability argument is relevant only in cases where EBIs are part of some unavoidable activity, and when the same cannot be said for NBIs. And as for physical invasiveness, I have myself testified to its limits to bear the full burden of vindicating the intuition against *Similarity*. So are these grounds sufficiently strong to undermine *Similarity*? Taken together, I deem that they are. Call this the cluster view, which states that there are multiple moral grounds that, when considered together, undermine *Similarity*. The cluster view casts doubt that there is a single moral ground that provides a knock-down argument against *Similarity* and that bears the full burden of vindicating the anti-*Similarity* intuition. But when taken together, they establish palpable moral separation suggested by the intuition. When we consider physical invasiveness, disclosure and avoidance, and inevitability together in the cases with which I've engaged, we reinforce the sense that Cafeteria B intervenes more problematically than Cafeteria A in *Cafeterias*, or that Physician B acts more objectionably towards her patient than Physician A in *Examination rooms*. Note however that not all considerations seem as important in the other two pairs of cases. Disclosure seems a less strong consideration in *City policies* than in the previous two pairs, given that the EBI doesn't target one particular decision but behavior more generally. Both disclosure and inevitability hardly justify the EBI of Diner A in *Diners*, seeing that the EBI doesn't seem to afford avoidance, or that while painting walls may be part of standard Diner activities, it isn't the case that walls must inevitably induce hunger. Further moral grounds may be required to reinforce anti-*Similarity* in these two latter cases.

Endorsing the cluster view also comes at a cost; it has a hard time explaining the intuition, since the intuition is unlikely derived from various intricate moral sources. Nevertheless, the task here was not to give a full explanation of the intuition, but only to vindicate it. The cluster view in the version that I offer here is not meant to be the final word on the various moral grounds that separate EBIs and NBIs; more grounds may be added along the way. But it is sufficient, in my view, for establishing discernible separation.

Acknowledgements For insightful comments and suggestions on various versions of the paper, I would like to thank Tom Douglas, Gabriel De Marco, Nino Kadić, Lovro Savić, Aleksandar Simić, Zlata Božac, Andrés Moles, two anonymous reviewers, and the Ethics of Behavioural Influence and Prediction Work-in-Progress Group, as well as the audience at 'SINe Neuroethics in a Time of Global Crises' (Milano, May 2022).

Declarations

Conflict of Interest The author declares that he has no conflict of interest.

References

1. Bleicher, S. 2005. *Contemporary color: Theory and use*. Clifton Park: Thomson/Delmar Learning.
2. Berman, M. 2007. *Street-smart advertising: How to win the battle of the buzz*. Lanham: Rowman & Littlefield.
3. Wren, A.M., L.J. Seal, M.A. Cohen, A.E. Brynes, G.S. Frost, K.G. Murphy, W.S. Dhillo, and S.R. Bloom. 2001. Ghrelin enhances appetite and increases food intake in humans. *The Journal of Clinical Endocrinology & Metabolism* 86 (12): 5992–5995. https://doi.org/10.1210/jcem.86.12.8111.
4. Bublitz, J., and R. Merkel. 2014. Crimes against minds: On mental manipulations, harms and a human right to mental self-determination. *Criminal Law and Philosophy* 8 (1): 51–77. https://doi.org/10.1007/s11572-012-9172-y.
5. Marcano-Olivier, M., Pearson, R., Ruparell, A., Horne, O.J., Viktor, S., Erjavec, M. 2019. A low-cost Behavioural Nudge and choice architecture intervention targeting school lunches increases children's consumption of fruit: a cluster randomized trial. *International Journal of Behavioral Nutrition and Physical Activity* 16. https://doi.org/10.1186/s12966-019-0773-x.
6. Douglas, T. 2022. The mere substitution defence of nudging works for neurointerventions too. *Journal of Applied Philosophy* 39 (3): 407–420. https://doi.org/10.1111/japp.12568.
7. Rehman, S.U., P.J. Nietert, D.W. Cope, and A.O. Kilpatrick. 2005. What to wear today? Effect of doctor's attire on the trust and confidence of patients. *American Journal*

of Medicine 118 (11): 1279–1286. https://doi.org/10.1016/j.amjmed.2005.04.026.
8. Gelfand, S.D. 2016. The meta-nudge – a response to the claim that the use of nudges during the informed consent process is unavoidable. *Bioethics* 30 (8): 601–608. https://doi.org/10.1111/bioe.12266.
9. Helbich, M., D. de Beurs, M.-P. Kwan, R.C. O'Connor, and P.P. Groenewegen. 2018. Natural environments and suicide mortality in the Netherlands: A cross-sectional, ecological study. *Lancet Planetary Health* 2 (3): 134–139. https://doi.org/10.1016/S2542-5196(18)30033-0.
10. Schrauzer, G.N., and K.P. Shrestha. 1990. Lithium in drinking water and the incidences of crimes, suicides, and arrests related to drug addictions. *Biological Trace Element Research* 25 (2): 105–113. https://doi.org/10.1007/BF02990271.
11. Vita, A., L. De Peri, and E. Sacchetti. 2015. Lithium in drinking water and suicide prevention: A review of the evidence. *International Clinical Psychopharmacology* 30 (1): 1–5. https://doi.org/10.1097/YIC.0000000000000048.
12. Thaler, R.H., and C.R. Sunstein. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. New Haven: Yale University Press.
13. Gilovich, T., D. Griffin, and D. Kahneman, eds. 2002. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge University Press.
14. Sunstein, C.R. 2014. *Why nudge? The politics of libertarian paternalism*. New Haven & London: Yale University Press.
15. Krishnamurthy, M. 2015. Nudging global poverty alleviation? *The Law & Ethics of Human Rights* 9 (2): 249–264. https://doi.org/10.1515/lehr-2015-0008.
16. Beraldo, S., and J. Karpus. 2021. Nudging to donate organs: Do what you like or like what we do? *Medicine, Health Care and Philosophy* 24 (3): 329–340. https://doi.org/10.1007/s11019-021-10007-6.
17. Moles, A. 2015. Nudging for liberals. *Social Theory and Practice* 41 (4): 644–667. https://doi.org/10.5840/soctheorpract201541435.
18. Grüne-Yanoff, T. 2012. Old wine in new casks: Libertarian paternalism still violates liberal principles. *Social Choice and Welfare* 38 (4): 635–645. https://doi.org/10.1007/s00355-011-0636-0.
19. Rozeboom, G. 2020. Nudging for rationality and self-governance. *Ethics* 131 (1): 107–121. https://doi.org/10.1086/709986.
20. Kosfeld, M., M. Heinrichs, P.J. Zak, U. Fischbacher, and E. Fehr. 2005. Oxytocin increases trust in humans. *Nature* 435 (7042): 673–676. https://doi.org/10.1038/nature03701.
21. Zak, P.J., R. Kurzban, and W.T. Matzner. 2004. The neurobiology of trust. *Annals of the New York Academy of Sciences* 1032: 224–227. https://doi.org/10.1196/annals.1314.025.
22. Tse, W.S., and A.J. Bond. 2002. Serotonergic intervention affects both social dominance and affiliative behaviour. *Psychopharmacology* 161 (3): 324–330. https://doi.org/10.1007/s00213-002-1049-7.
23. Douglas, T. 2015. The morality of moral neuroenhancement. In *Handbook of neuroethics*, ed. J. Clausen and N. Levy, 1227–1249. Dordrecht: Springer.
24. Persson, I., and J. Savulescu. 2012. *Unfit for the future: The need for moral enhancement*. Oxford: Oxford University Press.
25. Crutchfield, P. 2019. Compulsory moral bioenhancement should be covert. *Bioethics* 33 (1): 112–121. https://doi.org/10.1111/bioe.12496.
26. Harris, J. 2011. Moral enhancement and freedom. *Bioethics* 25 (2): 102–111. https://doi.org/10.1111/j.1467-8519.2010.01854.x.
27. Rakić, V. 2014. Voluntary moral enhancement and the survival-at-any-cost bias. *Journal of Medical Ethics* 40 (4): 246–250. https://doi.org/10.1136/medethics-2012-100700.
28. Fischer, J.M., and M. Ravizza. 1998. *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.
29. Pettit, P. 1997. *Republicanism: A theory of freedom and government*. Oxford: Clarendon Press.
30. Sparrow, R. 2014. Better living through chemistry? A reply to Savulescu and Persson on "moral enhancement." *Journal of Applied Philosophy* 31 (1): 23–32. https://doi.org/10.1111/japp.12038.
31. Selinger, E., and K.P. Whyte. 2012. Nudging cannot solve complex policy problems. *European Journal of Risk Regulation* 3 (1): 26–31. https://doi.org/10.1017/S1867299X0000177X.
32. John, P., G. Smith, and G. Stoker. 2009. Nudge nudge, think think: Two strategies for changing civic behaviour. *The Political Quarterly* 80 (3): 361–370. https://doi.org/10.1111/j.1467-923X.2009.02001.x.
33. Schaefer, G.O. 2015. Direct vs. indirect moral enhancement. *Kennedy Institute of Ethics Journal* 25 (3): 261–289. https://doi.org/10.1353/ken.2015.0016.
34. Levy, N. 2007. *Neuroethics*. Cambridge: Cambridge University Press.
35. Douglas, T. 2018. Neural and environmental modulation of motivation: what's the moral difference? In *Treatment for crime: Philosophical essays on neurointerventions in criminal justice*, ed. D. Birks and T. Douglas, 208–223. Oxford: Oxford University Press.
36. Douglas, T. 2022. If nudges treat their targets as rational agents, nonconsensual neurointerventions can too. *Ethical Theory and Moral Practice* 25 (3): 369–384. https://doi.org/10.1007/s10677-022-10285-w.
37. Pugh, J. 2019. Moral bio-enhancement, freedom, value and the parity principle. *Topoi* 38 (1): 73–86. https://doi.org/10.1007/s11245-017-9482-8.
38. Fitzpatrick, W. 2006. The intend/foresee distinction and the problem of "closeness." *Philosophical Studies* 128 (3): 585–617. https://doi.org/10.1007/s11098-004-7824-z.
39. Morrison, M., S. Gan, C. Dubelaar, and H. Oppewal. 2011. In-store music and aroma influences on shopper behavior and satisfaction. *Journal of Business Research* 64 (6): 558–564. https://doi.org/10.1016/j.jbusres.2010.06.006.
40. Douglas, T. 2013. Moral enhancement via direct emotion modulation: A reply to John Harris. *Bioethics* 27 (3): 160–168. https://doi.org/10.1111/j.1467-8519.2011.01919.x.
41 Bovens, L. 2009. The ethics of nudge. In *Preference change: Approaches from philosophy, economics and*

psychology, ed. T. Grüne-Yanoff and S.O. Hansson, 207–219. Berlin & New Yor: Springer.
42. Coons, C., and M. Weber. 2013. Introduction: Paternalism – issues and trends. In *Paternalism: Theory and practice*, ed. C. Coons and M. Weber, 1–24. New York: Cambridge University Press.
43. Loewenstein, G., C. Bryce, D. Hagmann, and S. Rajpal. 2015. Warning: You are about to be nudged. *Behavioral Science and Policy* 1: 35–42.
44. Ivanković, V., and B. Engelen. 2019. Nudging, transparency, and watchfulness. *Social Theory and Practice* 45 (1): 43–73. https://doi.org/10.5840/soctheorpract20191751.
45. Miller, P.M., and N.S. Fagley. 1991. The effects of framing, problem variations, and providing rationale on choice. *Personality and Social Psychology Bulletin* 17 (5): 517–522. https://doi.org/10.1177/0146167291175006.
46. Sieck, W., and J.F. Yates. 1997. Exposition effects on decision making: choice and confidence in choice. *Organizational Behavior and Human Decision Processes* 70 (3): 207–219. https://doi.org/10.1006/obhd.1997.2706.
47. Mitchell, G. 2005. Libertarian paternalism is an oxymoron. *Northwestern University Law Review* 99 (3): 1245–1278.
48. Almashat, S., B. Ayotte, B. Edelstein, and J. Margrett. 2008. Framing effect debiasing in medical decision making. *Patient Education and Counseling* 71 (1): 102–107. https://doi.org/10.1016/j.pec.2007.11.004.
49. Grill, K. 2014. Expanding the nudge: Designing choice contexts and choice contents. *Rationality, Markets and Morals* 5 (90): 139–162.
50. White, M. 2013. *The manipulation of choice: Ethics and libertarian paternalism*. New York: Palgrave Macmillan.
51. Hausman, D.M., and B. Welch. 2010. Debate: To nudge or not to nudge. *Journal of Political Philosophy* 18 (1): 123–136. https://doi.org/10.1111/j.1467-9760.2009.00351.x.
52. Sunstein, C.R. 2015. Nudges, agency, and abstraction: A reply to critics. *Review of Philosophy and Psychology* 6 (3): 511–529. https://doi.org/10.1007/s13164-015-0266-z.
53. Sunstein, C.R. 2016. The ethics of choice architecture. In *Choice architecture in democracies: Exploring the legitimacy of nudging*, ed. A. Kemmerer, C. Möllers, M. Steinbeis, and G. Wagner, 21–74. Baden-Baden: Nomos Verlagsgesellschaft.
54. Cohen, S. 2013. Nudging and informed consent. *American Journal of Bioethics* 13 (6): 3–11. https://doi.org/10.1080/15265161.2013.781704.
55. Brooks, T. 2013. Should we nudge informed consent. *American Journal of Bioethics* 13 (6): 22–23. https://doi.org/10.1080/15265161.2013.781710.
56. Engelen, B., A. Thomas, A. Archer, and N. van de Ven. 2018. Exemplars and nudges: Combining two strategies for moral education. *Journal of Moral Education* 47 (3): 346–365. https://doi.org/10.1080/03057240.2017.1396966.
57. Blumenthal-Barby, J.S. 2013. Choice architecture: A mechanism for improving decisions while preserving liberty? In *Paternalism: Theory and practice*, ed. C. Coons and M. Weber, 178–196. New York: Cambridge University Press.
58. Engelen, B. 2019. Ethical criteria for health-promoting nudges: A case-by-case analysis. *American Journal of Bioethics* 19 (5): 48–59. https://doi.org/10.1080/15265161.2019.1588411.
59. Blumenthal-Barby, J.S. 2021. *Good ethics and bad choices: The relevance of behavioral economics for medical ethics*. Cambridge, MA: The MIT Press.
60. Buchanan, A., and R. Powell. 2018. *The evolution of moral progress: A biocultural theory*. Oxford: Oxford University Press.
61. Melo-Martin, I., and A. Salles. 2015. Moral bioenhncement: Much ado about nothing? *Bioethics* 29 (4): 223–232. https://doi.org/10.1111/bioe.12100.
62. Wiseman, H. 2014. SSRIs as moral enhancement interventions: A practical dead end. *AJOB Neuroscience* 5 (3): 21–30. https://doi.org/10.1080/21507740.2014.911214.
63. Levy N. 2019. Nudge, nudge, wink, wink: Nudging is giving reasons. *Ergo 6* (10). https://doi.org/10.3998/ergo.12405314.0006.010
64. DeGrazia, D. 2014. Moral enhancement, freedom, and what we (should) value in moral behaviour. *Journal of Medical Ethics* 40 (6): 361–368. https://doi.org/10.1136/medethics-2012-101157.