Internal and external moral enhancements: the ethical parity principle and the case for a prioritization

Matteo Galletti

Abstract: Is there any moral difference between internal moral enhancements, which directly affect the biological nature of human beings, and external moral enhancements, which nudge choices and behavior without changing human biology? If Neil Levy's Ethical Parity Principle is applied, the answer should be no. Recently, John Danaher has argued that the Ethical Parity Principle is invalid and that there are ethical and political reasons for a prioritization of internal over external moral enhancements. Although Danaher's argument presents some interesting insights, it needs to be corrected with finer-grained distinctions of the types of moral enhancements.

Keywords: moral enhancement, behavioral ethics, nudge, procedural moral enhancement, ethical parity principle.

1. *Moral Enhancements: Internal and External*

In the debate on moral enhancement, one of the proposed but little-discussed distinctions is that between internal and external enhancements. Before introducing it, however, I need to clarify what moral enhancement means. Following some suggestions from DeGrazia (2013), I propose this definition: an enhancement is any deliberate intervention that strengthens or reduces existing capacities and dispositions or creates new ones to improve the motivation, decision-making, and behavior of an individual or population in the moral domain.

This definition raises some questions. First, it includes under the label of moral enhancement interventions that do not increase moral traits and capacities but mitigate or eliminate certain tendencies deemed to be pernicious, such as dispositions to violent reaction, implicit biases, or, generically, "counter-moral" emotions (Douglas 2008). Suppose an effect is to improve an individual's moral condition. In that case, we consider it indifferent whether the way by which it is achieved is active (increase or reinforcement) or negative (erasure or reduction).

Second, according to some authors, a moral enhancement qualifies as such because of its effect, regardless of the intentions with which it is implemented. I find this objection reasonable, but the reference to intentionality allows

philinq XI, 2-2023, pp. 57-72
ISSN (print) 2281-8618-ETS doi: 10.4454/philinq.v11i2.492

for a distinction between "enhancement" and "improvement", which I think is relevant to the judgment of this kind of technique. An "improvement" occurs when a given intervention (increase or decrease) on some property of the organism betters its condition so that at instant t0 (prior to the enhancement), the condition of the organism is X, and at instant t1 (after the enhancement) the condition of the organism is Y, where Y is judged better than X. An enhancement aimed at improving could be causally effective in reducing or increasing a certain capacity but not result in an actual improvement in the individual's condition: one might think that an enhancement that endows an individual with the stature of 3 meters would provide the individual with a positional good because he or she will have an advantage in some activities (e.g., sports), but might adversely affect many other areas of his or her life, given the difficult adaptability of his or her stature to the surrounding environment. Alternatively, the enhancer, aware of this outcome, might practice such an intervention precisely with the intention of harming the individual. The distinction between "enhancement" and "improvement" allows finer-grained judgments.

Third, the triad being addressed (judgment, motivation, behavior) includes some rather heterogeneous elements of morality (judgment and motivation belong to psychology, and behavior indicates an observable external expression). Nevertheless, the central distinction here is between the term "dispositions", which includes character traits and moral dispositions such as altruism or empathy, and the concept of "capacities", which refers to second-order reflective capacities, such as moral reasoning, deliberation, and imagination. I will return to this distinction later because it is revealing for understanding the moral status of various types of enhancement.

If we accept this general definition of "moral enhancement", an internal enhancement is a deliberate intervention that either strengthens or reduces existing capacities and dispositions or creates new ones by acting directly on biology, with the aim of improving the motivation, decision-making, and behavior of an individual or population in the moral domain. For example, drug administration and genome editing intervention on the somatic or germline that have this effect can be considered internal enhancements. Thus, internal enhancements involve integrating the biotechnological intervention into the organism's biology. I propose to reserve the name "moral bio-enhancements" (MBE) for these interventions.

External enhancements consist of other means of improving moral traits and abilities, such as external devices, without directly affecting or integrating with the organism's biological constitution; some other external enhancements introduce changes in the context to achieve the enhancing effect. Examples of external enhancements are the use of artificial intelligence devices to make moral

decisions (Borenstein *et al.* 2016; Giubilini *et al.* 2018; Lara 2021) or specific changes in the context of choice that affect sensory perception and favor certain moral judgments (Schnall *et al.* 2008; Wheatley *et al.* 2005; Eskine *et al.* 2011). I propose to call these interventions "moral environment-enhancements" (MEE).

As I understand it, the distinction between MBE and MEE does not perfectly overlap with the distinction between biotechnological enhancements and so-called "traditional" enhancements, such as education, socialization, and the organization of a system of rewards and punishments. Traditional enhancements can, at best, be considered as a species of the MEE genre, including advanced technological interventions such as AI devices or, as we shall see, choice architectures. It is precisely the behavioral sciences that are helping to provide a description of moral agency whose natural limits necessitate the use of innovative measures to influence ex-ante the actions of individuals. In this essay, we will focus on a particular type of these external enhancements, namely so-called "nudges", to induce morally desirable behavior in individuals.

The literature on moral enhancement has only occasionally considered the distinction between external and internal moral enhancements. In this essay, we will argue that any judgment of the moral superiority of MBEs over MEEs, such as moral nudges, cannot be general in nature but must be circumstantial. We will consider Danaher's recent contribution to the debate.

2. *Behavioral Ethics and Moral Nudges*

The numerous empirical research in psychology and behavioral economics have defined a field of studies, which includes rather heterogeneous approaches to the phenomenon of morality, called "behavioral ethics" (BE) (Bazerman *et al.* 2012). BE "addresses people's inability to fully recognize the ethical, moral and legal aspects of their behavior" (Feldman 2018: 2); although BE shares with behavioral analysis the empirically supported belief that biases affect individual choices, it departs from it concerning the general explanation of how these biases work. According to the behavioral sciences, biases are due to the involvement of automatic responses, unmediated by reflexivity and deliberative reasoning, that take the form of post hoc rationalization to justify unethical behavior. Instead, BE provides a more complex picture of moral agency, in which the limitations of various cognitive capacities are due to the tendency to seek self-interest and the inherent need to maintain a coherent and positive self-representation. The situation in which the agent chooses also has a limiting impact on perception, judgment, and choice. The action of these mechanisms occurs mostly unconsciously and may also resort to post-hoc strategies of moral disengagement, thus leading to a hiatus between full personal awareness of what the

agent is doing and the actual intention to do harm. In general, ethical biases inhibit individuals' ability to recognize the moral quality of their actions (35-36).

Innovative tools for intervening in human behavior (88-98) include nudges, which can be defined as "ways of influencing choice without limiting the choice set or making alternatives appreciably more costly in terms of time, trouble, social sanctions, and so forth. They are called for because of flaws in individual decision-making, and they work by using those flaws" (Hausman *et al.* 2010, 124; Mongin *et al.* 2018).

The activity of nudging is based on the division of the mind's architecture into fast, parallel, automatic, associative, effortless psychological processes ("System 1") and slow, serial, controlled processes that require effort and are governed by rules ("System 2"). Based on this bipartition, one can then distinguish the nudges that exploit unintentional processes in System 1 from the nudges that instead enhance agents' reflexive and self-control abilities. For example, thrusts of the first type are interventions that exploit decision inertia and bias toward the status quo, whereby people tend not to make choices other than those they are accustomed to, or not to change the given situation, even when a change in the status quo could be beneficial. Choice architects can intervene by setting the most rational or most beneficial option, predicting that the agent will most likely tend not to change it. Other examples include exploiting the framing effect, that is, the disposition of agents to have different reactions to the same information when it is phrased in different ways, and the use of explicit imagery to make certain information more salient (think of the design of cigarette packages to make smokers more aware of the harms of smoking). Second-type nudges enable individuals to translate their intention into actual choices and actions and to avoid falling into the traps of the weakness of will. Agents can better understand information regarding specific products or situations to make more rational choices. "Educational nudges" do not exploit cognitive or decision-making limitations but enhance deliberative and executive skills. Generally, nudges do not coerce people to choose and act in a certain way, but they "guide" behavior, allowing agents to choose and act otherwise.

Nudges can be deployed in a paternalistic framework, when the choice architecture guides individuals' choices for the purposes of increasing their welfare, or in a non-paternalistic framework, when nudges guide individuals' choices to produce more externalities. The distinction is unclear, however, because deficient individual choices can also create negative externalities (Carlsson *et al.* 2021: 216), but I only consider non-paternalistic nudges since my focus is on BE. Nudges of this kind can be employed to induce actions respectful of significant community goods, such as reducing resource consumption or adopting ecologically compatible conduct (Carlsson *et al.* 2021; Wee *et al.* 2021; Santos

Silva 2022), or for their generic moral effect, insofar they increase altruistic and generally prosocial actions (Gråd *et al.* 2024; Valerio *et al.* 2021; Dimant *et al.* 2022). We will call these kinds of behavioral interventions "moral nudges".

Like proponents of behavioral ethics, the advocates of internal moral enhancements recognize the limitations of human nature as well: Persson and Savulescu (2012), for example, accurately describe human moral psychology as an evolutionary product adapted to very different environmental challenges than the current global ones and list many biases that need to be corrected. Walker claims that we can try to answer the question of why evil actions happen with great frequency by invoking views about humans' "defective natures", or the fact that "humans are innately evil is" (Walker 2009: 28-29). However, some proponents of MBE are skeptical about the efficacy of BE tools. For example, Persson and Savulescu note that nudges should be easy to avoid so that agents are genuinely free to choose otherwise; for this reason, behavioral tools such as nudges "are better suited to make us overcome backsliding on isolated occasions or to make us execute what we already think is best for us, or to make us decide between roughly equally balanced alternatives" (Persson *et al.* 2012: 79, footnote 2). The problem then lies in the effectiveness of these tools over time, which is not related to the ethical issues raised by nudging.

3. *The Automation Problem*

Recently, John Danaher (2019) argued for MEE's moral and political primacy to MBE. Danaher takes a broad definition of "moral enhancement": all interventions that enhance human moral judgment and behavior fall into this category, which includes "anything that develops morally-relevant emotions (such as trust or empathy), or virtues (such as courage and generosity), morally-relevant reasoning capacities (such as evidential assessment, impartiality and lack of prejudice), or improves individual moral actions (such as helping and caring for others)" (40) and distinguishes between internal (BME) and external moral enhancement (MEE). Danaher considers the effects on individual capabilities of a specific type of MEE, namely the use of electronic or AI devices that drive us towards the desired goal. For example, smartphone apps that nudge individuals toward money donations for charitable associations or other moral purposes, or a bracelet that gives you an electric shock when you do something morally or behaviorally inadequate (41-42). But he adds that these devices imply a philosophy of nudge, "which is influential in the design of many of the contemporary behavior change policies, apps, and devices" (41, 49). So, we can generalize his conclusion to embrace all behavioral interventions, even those which do not resort to smart devices and intelligent algorithms.

Danaher's central thesis is that the asymmetry between external and external enhancements should lead us to evaluate the former as morally more acceptable than the latter. Such an asymmetry implies rejecting the Ethical Parity Principle (in its weakest form) formulated by Neil Levy. In its weak version, the Ethical Parity Principle (EPP) holds that if we find compelling reasons for considering morally problematic interventions to modify the external environment, we must apply the same reasons to internal interventions. Unlike the strong version, the weak version of the EPP does not require us to accept the ontological thesis of the extended mind; that is, we need not assume that the domain of the mental extends outside the heads of individuals to include aspects of the external environment as well (Levy 2007: 60-64). According to Danaher, Levy's Ethical Parity Principle is undermined by three salient moral differences between internal processes and external devices (*memory integration*, *fungibility*, and *consciousness*). Memory is dynamic, while information stored in a notebook o in a mobile phone is static; a destroyed notebook can be easily replaced, and the user can form new memories or store new information if she lost pictures and files, while someone who got her hippocampus destroyed has a permanent disability in creating new long-term memories; finally, internal functioning is more integrated into conscious experience than the functioning of external props, and the same goes about internal memories and memories stored in external devices.

The differences identified by Danaher are relative to the *nature* of internal and external processes, functioning, and content, and this "ontological" divide has practical implications: internal devices produce *internal automaticity*, defined as "the control of behavior by not-immediately-conscious neural networks" (Danaher 2019: 49), while external devices initiate *external automation* processes. External automation has effects that hardly integrate with our perception and understanding of the world because it can easily bypass our conscious moral reasoning. On the contrary, internal automaticity does not undermine the deliberative process. So, what is troubling with MEE is its impact on what we can call "reflective capacities". Actually, Danaher does not use this term. Still, I prefer to speak of "reflective capacities" instead of "conscious moral reasoning" in order not to take a markedly rationalistic position in metaethics and moral psychology. Reflective capacities can also be compatible with a sentimentalist and deliberative conception of ethics, thus leaving open the metaethical question of how to precisely define these reflective powers.

According to Danaher (47), a problem of political legitimacy arises here: for a political decision to be legitimate, it must follow reliable procedures that exhibit outcome-independent virtues, as well as produce predictably desirable consequences. How policymakers intend to achieve a specific objective is *also* meaningful from this perspective. The proceduralist view introduces the need

to respect values such as transparency, participativeness, and comprehensibility. According to Danaher, external enhancements are incompatible with the proceduralist idea because they threaten these values, violating the central commitment of liberal democratic democracies, i.e., the commitment to treating citizens as moral *agents*, as subjects capable of actively relating to the moral problems they encounter in their lives (including political challenges). Bypassing reflective capacities, they turn targets into *passive* recipients, which are manipulated to have a particular desirable output (behavior that becomes more altruistic, more generous, more just, etc., thus conforming to specific moral standards). Danaher concludes that internal enhancements are preferable to the external enhancements.

Before analyzing Danaher's proposal, I would like to point out that, appearances to the contrary, this approach seems in line with at least one of Levy's remarks in introducing the EPP. He claims that a mere difference between internal and external cannot lead to a refutation of the externality thesis, but it should be taken as a confirmation. The point is that external props are attractive to the extent that they succeed in securing a more conspicuous cognitive, emotive, or motivational gain than that achieved by internal processes. For this reason, we can find attractive the ontological hypothesis of the extended mind (Levy 2007: 59-60). However, Danaher draws a radically different conclusion from this "pragmatic" approach: the fact that external tools are inefficient or even harmful when used to morally enhance people because they turn agents into passive recipients is a reason to oppose them, but the same reason does not apply to internal modifications. The automaticity produced by internal modification is more acceptable than automation, and so the EEP is unsound. Allhoff, Lin, and Steinberg (2011) argue an analogous line. They argue that the spatial location of the enhancement does not entail any moral difference because there is no reason to believe that incorporation is morally questionable. The example they chose is perfectly in line with Levy's equality principle. There is no difference *in kind* between "a neural implant [that] gives access to Google and the rest of the online world [and] using a laptop computer or Pocket PC to access the same" (204). The embedding nature of the former is not diriment to the extent that both carry "the same capabilities with us" (204). Although there is no moral difference in *kind*, there may be one in the *outcome*. According to Allhoff, Lin, and Steinberg, the moral symmetry between the two enhancements is assured if they are both effective in securing a particular capability; so, they suggest that we have to look for potential moral differences, not in the way we enhance an organism, but in the effects and in the impact on human capabilities of the enhancement we employ.

4. *The Automation Problem and the Role of Relational Freedom*

I strongly agree with Danaher on two points: (1) it's not the external or internal nature of an intervention that makes the moral difference, but its effects on agency; (2) automation poses a problem of political legitimacy. In fact, I would go even further with the latter: automation establishes a certain political relationship between the agent and the recipient of the intervention. Take some political authority that uses a BE tool to push people toward more altruistic choices; this intervention establishes a not-reciprocal relationship between the nudger and the nudgee.

A reciprocal relationship presupposes some basic expectations on the part of the people interacting, partially modeled on certain standards internal to that relationship. Thus, it is part of the regular interaction between human beings to expect that no one will be harmed by the other without a valid reason for doing so and that specific adverse reactions to a violation of this expectation are appropriate because the wrongdoer has betrayed the minimum threshold of trust that characterizes the relationship. Of course, the degree of trust or suspicion we may have toward others also depends on the context; situations that are less secure or in which we have little information may motivate a cautious attitude, just as interactions with people with whom we are more familiar may change the nature of mutual expectations. Trust and the reactions that follow its intentional violation form the core of a very specific type of relationship, one between people capable of reciprocity, that is precluded when dealing with very young children or people with severe mental illness. The interaction that takes place between moral agents presupposes the adoption of a dual attitude: a normative expectation of others' behavior and a willingness to treat the other as a "participant" in a reciprocal relationship. When this twofold attitude is not possible, relationships are marked by less reciprocity, to the extreme end of the spectrum where an "objective" attitude prevails. In such a case, the other is no longer considered a responsible person but someone to be "cared for", "managed", or "directed" (Strawson 1962). The shift from the participative standpoint to the objective one characterizes the nudge intervention. Nudgers suspend participatory attitudes and adopt an objectifying perspective toward the nudgee. She is a passive recipient, or at least she is treated as such.

One can reply that the nudger has a valid reason to adopt an objective attitude because the nudgee is in a condition of moral deficiency similar to that of an ill person. However, this kind of generalization fails to consider that the moral quality of behavioral intervention also hinges on the type of relationship we expect between those who possess political authority and those who are the recipients of policies. The endorsement of an objective attitude on the part

of nudgers depends crucially on the relationship between citizens and public decision-makers and the mutual expectations that structure these relationships. In this broader context, citizens are not agents to be respected but patients to be managed. As Hausman and Welch (2010: 134) put it:

> If a government is supposed to treat its citizens as agents who, within the limits that derive from the rights and interests of others, determine the direction of their own lives, then it should be reluctant to use means to influence them other than rational persuasion. Even if, as seems to us obviously the case, the decision-making abilities of citizens are flawed and might not be significantly diminished by concerted efforts to exploit these flaws, an organized effort to shape choices still appears to be a form of disrespectful social control.

Even if the use of moral nudges does not harm people's autonomy and well-being, it still impacts on the expectations that citizens in liberal democracies may have of those who govern them. The question of whether behavior management conflicts with such expectations is not merely empirical because it concerns the background against which interpersonal relationships are given; the introduction of nudges alters this background without this transformation being subject to reflection and consideration. This objection can also be framed in terms of respect for decisional autonomy (Rebonato 2012: 200-207). Still, regardless of the normative language it expresses, it echoes some criticism of internal moral enhancements. For example, Robert Sparrow (2014) pointed out a fundamental disanalogy between moral enhancement through traditional ways, such as education, and MBE: traditional means establish a relationship of equality and respect between the enhancing subject and the enhanced one, which responds to norms that justify educational activity and are in principle acceptable to all involved in the enterprise.

On the other hand, biomedical interventions create an entirely different relationship because they "operate in an instrumental or technical mode" (Sparrow 2014: 26) that treats the enhanced as an object and not as a subject. Sparrow echoes Philipp Pettit's idea of freedom as the absence of domination. In fact, the imposition of an MBE puts the enhanced person in a condition of subjugation and deprives her of her status as a responsible agent. Similarly, Michael Hauskeller (2017: 373-374) claims that if X makes it psychologically impossible for Y to want to do anything other than what X desires, then X's control over Y is total. Employing MBEs assumes an objectifying attitude: it expresses a suspension of all participatory perspectives and induces one to regard those who behave unjustly not as moral agents harboring inadequate moral dispositions or feelings while remaining fully participants in the practices of moral responsibility but as objects to be manipulated and corrected.

Thus, the same kind of criticism has been leveled at MBEs that target dispositions and emotions, as well as at MEEs that exploit automatic bias and heuristics. In both cases, the recipient of the intervention performs specific actions because of an automation mechanism: when this state of affairs is the intentional product of an institution or another person, a relationship of domination and control is established, and it is incompatible with the recipient's relational freedom. Moreover, MBEs and MEEs are also on a par regarding outsourcing moral reasoning. Danaher (2019: 50) stresses that external enhancements outsource the reasoning process, relieving agents of a cognitive burden, but the same issue seems to affect both MBEs and MEEs. In the case of internal enhancements, it is biochemical functioning that causes the output, bypassing reflective processes. If X receives a drug that amplifies his sense of justice in negotiations, her choices will automatically be more just without any reflective activity on his part. As Danaher claims, "We don't need to think for ourselves; we don't need to weigh the moral reasons for and against a particular action; the algorithm does all that for us" (50). If we substitute "algorithm" with "the molecular action", we have a mirrored criticism of MBEs.

Danaher has another arrow in his quiver: internal automaticity can easily be integrated with the individual mindset. A reiterated use of MBEs can be transformed into a permanent disposition in the long run and incorporated into one's moral character. For instance, a bioenhanced judge may take more empathic decisions without being aware of "the immediate proximate cause of his or her decision to choose the morally superior outcome, but he or she may over time generate a more empathetic disposition, which will affect future interactions with the world, and will, over time, result in enhanced moral sensitivity and awareness" (49). Even external enhancements rely on non-transparent mechanisms which the subject is not aware of. Still, they cannot be integrated in this way and can have corrosive effects in the long run (for a different view, see Agar 2014: 46-47).

However, even in this case, there is no real difference. It may be that the empathic disposition fits in the individual moral character in a more spontaneous way than the prolonged effect of an external factor can do; the judge may progressively endorse the effect of BME on dispositions. But note that if the judge does not voluntarily choose to undergo BME or has not had a pro-attitude toward it, his case looks very much like the manipulation of his moral character over time without giving any consent to these changes. In the case of a nudge, if some degree of transparency is assured, the agent is aware that his behavior can be directed in a certain way by an external prompt and willingly accepts the outcome of this conditioning. The automatic choice can become integrated into his identity. It seems that in both situations, the only relevant factor is the

degree of subjective awareness of the impact of MBEs or MEEs on judgment, choice, motivation, and behavior.

5. *Defeating the Ethical-Political Illegitimacy of Moral Enhancement*

To summarize, there are plausible reasons to deny that there is a morally relevant difference between BMEs and MEEs concerning the effects on moral agency. In both cases, the same kind of automation puts the enhanced or nudged individual under the control of who administers an enhancing drug or introduces a nudge in the choice context. In both cases, the same issue of political legitimacy arises. At this point, two conditions should avoid the concerns raised by automation. These conditions are defeaters that block the normative power of automation.

The first defeater is the presence of awareness or voluntary endorsement of the intervention. Alfano and Robichaud (2018: 242-244) see using (moral and non-moral) nudges as a responsibility-conferring practice. When institutions and private officers nudge individuals, they exercise power to attribute to the nudgees a forward-looking responsibility for distinct values. The values realized by nudges are varied; they claim that nudgers are more justified in using nudges when these tools induce individuals to fulfill obligations to themselves or others, while the power to resort to them is less supported when it is at stake the production of goods for self and others. A nudger has the ability to assign forward-looking responsibility for meeting some obligation via nudges only when the nudgee is liable for this assignment.

However, there are domains in which nudges are immune from the attribution of responsibility. For example, it seems morally unwarranted to nudge for sensitive choices such as marital decisions, voting, or healthcare decisions (although the introduction of nudges in the medical context is controversial and there are many proposals to use behavioral economics tools for clinical decisions or to obtain patients' informed consent more easily). Furthermore, they mention another possibility of being immune from nudge: individuals and communities can repudiate the responsibility assignment explicitly (i.e., through voting) or hypothetically (while it is unclear what form a hypothetical repudiation can take). Similarly, they can accept responsibility for values through an explicit or hypothetical endorsement, thus conferring political legitimacy to nudges (246). Alfano and Robichaud focus on the community and political levels. Still, an individual can reject a nudge if they are aware of its existence and operation, as I have already mentioned.

Further, citizens can become choice architects and opt for self-nudging (Reijula *et al.* 2022). Even in the case of BMEs, the agent can take a position by ac-

cepting or rejecting the intervention. This is the case of the so-called voluntary BME: one who freely chooses to bio-enhance herself shapes her future desires and intentions and expresses a solicitude for the moral quality of his future self. Voluntary BME is thus a strategy of preventive self-control, an "essential constraint" to use Jon Elster's terminology, i.e., a dodge by which agents self-impose restrictions to condition future behavior because of some expected benefit: voluntary BME expresses a "certain form of rationality over time" (Elster 2000).

The second defeating condition is related to a change in the target of biological or behavioral interventions. Such strategies can enhance reflective capacities instead of biases, emotions, and dispositions. The problem with voluntary and non-voluntary BMEs and MMEs is that they modify behavior, leaving the individual moral character untouched. They maximize good outcomes but do not correct moral flaws (Simkulet 2016). But we have other biological and behavioral ways to obtain real *moral improvement* in individuals.

Indeed, a possible alternative approach to nudging is the so-called "boosts", "interventions that make it easier for people to exercise their agency by fostering existing competencies or instilling new ones" (Hertwig *et al.* 2017: 974). The boosts approach shows significant differences from the nudge approach. First, it views agents not as passive recipients but as decision-makers "whose competencies can be improved by enriching [their] repertoire of skills and decision tools and/or by restructuring the environment such that existing skills and tools can be more effectively applied" (Grüne-Yanoff *et al.* 2016: 152). Second, it is interested not only in the *outcome* of decision-making (the conformity of behavior to specific standards of rationality and/or morality) but is concerned with the *process* through which such an outcome is achieved. Third, it does not demand to adapt the individual mind to the choice environment by exploiting its cognitive flaws in order to guide behavior but modifies the choice environment to suit the reflective powers of human beings. Fourth, its concern is in decision makers being aware of the limits of their minds and the errors they make in their judgments and decisions: the boosting approach requires the active cooperation of individuals (they are offers that can be accepted or declined).

Thus, the boosting approach aims to enhance subjects' cognitive, reflective, and deliberative features or, to use the language of the dual structure of the mind, they seek to educate System 1, by employing tools such as reminders, warnings, information labels, etc. The functioning of boosts is not dissimilar to the nudges that Sunstein calls "educational" or to other alternative approaches in BE as ethical debiasing, training, and moral disambiguation. Take, for example, the last one. In many situations, there is ambiguity about the existence of a conflict of interest, so that people tend to convince themselves of the absence

of the conflict and to excuse their immoral choices. A choice architecture that eliminates ambiguity can partly resolve the problem and mobilize individual moral resources to avoid immoral behavior (Feldman 2018: 98-100, 102-104).

Similarly, MBEs that succeed in directly amplifying or indirectly facilitating capacities of moral reflection seem to avoid automation. This is referred to in the literature as "procedural" or "indirect" MBE, which does not target specific moral dispositions, but enhances the capacity to correct feelings and instinctive reactions, drawing on a wide range of cognitive and noncognitive, individual, and social resources (Raus *et al.* 2014: 268-270; Schaefer 2015; Schaefer *et al.* 2019). Procedural MBEs do not guarantee effective behavior change, but they are, in principle, more respectful of individual moral agency. They allow the enhanced agents to make free choices and thus moral mistakes, hopefully learning from them. In addition, they allow "the enhancer to remain neutral on a wide range of substantive moral positions". Even if the enhancers "cannot be completely substantively neutral, […] the range and type of substantive issues within the scope of the enhancer are severely limited" (Schaefer 2015: 274). From a metaethical point of view, those who defend procedural MBEs cannot be neutral because they should take sides in the controversy between moral rationalism and moral sentimentalism. Nevertheless, the point is that enhancing moral deliberation, reasoning, and imagination can preserve moral agency from automation issues and the risk of being controlled by second parties.

Procedural MBEs and boosting MEEs can produce a moral *enhancement* that is simultaneously a real *moral improvement* because they make people more reflective in the moral domain without compelling them to make the morally correct choice.

6. *Conclusion*

Danaher's arguments against moral parity between MEEs and MBEs need to be more convincing because the risk of automation is substantial for both groups of interventions. But the distinction between automaticity and automation has heuristic value. It can serve as a basis for identifying finer-grained concepts for distinguishing enhancements that pose a problem of ethical-political legitimacy from enhancements that succeed instead in ensuring effective moral improvement. Moral boosts (or "educational nudges") and procedural or indirect forms of bio-enhancement fall into this second category and we have a moral reason to prioritize them over internal enhancements of moral dispositions and the use of nudges.

Finally, I would suggest that the internal or external location of ME interventions is not morally relevant; it is a different spatial metaphor that is morally

pertinent, namely whether the ME intervention is "high" because it targets individual reflective capacities, or "low" because it takes aim at automatic dispositions and behaviors.

Matteo Galletti
University of Florence
matteo.galletti@unifi.it

References

Agar, Nicholas, 2014, *Truly Human Enhancement: A Philosophical Defense of Limits*, The MIT Press, Cambridge.

Alfano, Mark, Philip Robichaud, 2018, "Nudges and Other Moral Technologies in the Context of Power: Assigning and Accepting Responsibility" in David Boonin, ed., *The Palgrave Handbook of Philosophy and Public Policy*, Palgrave Macmillan, Cham: 235-248.

Allhoff, Fritz, *et al.*, 2011, "Ethics of Human Enhancement: An Executive Summary", in *Science and Engineering Ethics*, 17, 2: 201-212.

Bazerman, Max *et al.*, 2012, "Behavioral Ethics: Toward a Deeper Understanding of Moral Judgment and Dishonesty" in *Annual Review of Law and Social Science*, 8: 85-104.

Borenstein, Jason *et al.*, 2016, "Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being" in *Science and Engineering Ethics*, 22, 1: 31-46.

Capraro, Valerio, Jagfeld, Glorianna, Klein, Rana *et al.*, 2019, "Increasing Altruistic and Cooperative Behaviour with Simple Moral Nudges" in *Scientific Reports* 9, art. N. 11880: 1-11.

Carlsson, Fredrik, *et al.*, 2021, "The Use of Green Nudges as an Environmental Policy Instrument", in *Review of Environmental Economics and Policy*, 15, 2: 216-237.

Danaher, John, 2019, "Why Internal Moral Enhancement Might Be Politically Better than External Moral Enhancement" in *Neuroethics*, 12, 1: 39-54.

DeGrazia, David, 2014, "Moral Enhancement, Freedom, and What We (Should) Value in Moral Behaviour" in *Journal of Medical Ethics*, 40, 6: 361-368.

Dimant Eugen *et al.*, 2022, "Meta-Nudging Honesty: Past, Present, and Future of the Research Frontier", in *Current Opinion in Psychology*, 47: 1-4.

Douglas, Thomas, 2008, "Moral Enhancement", in *Journal of Applied Philosophy*, 25, 3: 228-245.

Elster, Jon, 2000, *Ulysses Unbound. Studies in Rationality, Precommitment, and Constraints*, Cambridge University Press, Cambridge.

Eskine, Kendall J. *et al.*, 2011, "A Bad Taste in the Mouth: Gustatory Disgust Influences Moral Judgment", in *Psychological Science*, 22, 3: 295-299.

Feldman, Yuval, 2018, *The Law of Good People. Challenging States' Ability to Regulate Human Behavior*, Cambridge University Press, Cambridge.

Giubilini, Alberto *et al.*, 2018, "The Artificial Moral Advisor. The 'Ideal Observer' Meets Artificial Intelligence", in *Philosophy & Technology*, 31, 2: 169-188.

Gråd, Erik *et al.*, 2024, "Do Nudges Crowd Out Prosocial Behavior?", in *Behavioural Public Policy*, 8, 1: 107-120.

Grüne-Yanoff, Till *et al.*, 2016, "Nudge Versus Boost: How Coherent are Policy and Theory?", in *Minds & Machines*, 26, 1-2: 149-183.

Hauskeller, Michael, 2017, "Is It Desirable to Be Able to Do the Undesirable? Moral Bioenhancement and the Little Alex Problem" in *Cambridge Quarterly of Healthcare Ethics*, 26, 3: 365-376.

Hausman, Daniel M. *et al.*, 2010, "Debate: To Nudge or Not to Nudge", in *The Journal of Political Philosophy*, 18, 1: 123-136.

Hertwig, Ralph *et al.*, 2017, "Nudging and Boosting: Steering or Empowering Good Decisions" in *Perspectives on Psychological Science*, 12, 6: 973-986.

Lara, Francisco, 2021, "Why a Virtual Assistant for Moral Enhancement When We Could Have a Socrates?" in *Science and Engineering Ethics*, 27, 42: 1-27.

Levy, Neil, 2007, *Neuroethics: Challenges for the 21st Century*, Cambridge University Press, Cambridge.

Mongin, Philippe *et al.*, 2018, "Rethinking Nudge: Not One but Three Concepts" in *Behavioural Public Policy*, 2, 1: 107-124.

Persson, Ingmar *et al.*, 2012, *Unfit for the Future. The Need for Moral Enhancement*, Oxford University Press, Oxford.

Raus, Kasper *et al.*, 2014, "On Defining Moral Enhancement: A Clarificatory Taxonomy" in *Neuroethics*, 7: 263-273.

Rebonato, Riccardo, 2012, *Taking Liberties. A Critical Examination of Libertarian Paternalism*, Palgrave MacMillan, Basingstoke.

Reijula, Samuli *et al.*, 2002, "Self-nudging and the Citizen Choice Architect" in *Behavioural Public Policy*, 6, 1: 119-149.

Santos Silva, Marta, 2022, "Nudging and Other Behaviourally Based Policies as Enablers for Environmental Sustainability" in *Laws*, 11, 9: 1-13.

Schaefer, G. Owen, 2015, "Direct vs. Indirect Moral Enhancement" in *Kennedy Institute of Ethics Journal*, 25, 3: 261-289.

Schaefer, G. Owen *et al.*, 2019, "Procedural Moral Enhancement" in *Neuroethics*, 12: 73-84.

Schnall, Simone, *et al.*, 2008, "With a Clean Conscience: Cleanliness Reduces the Severity of Moral Judgments" in *Psychological Science*, 19, 12: 2008: 1219-1222.

Simkulet, William, 2016, "Intention and Moral Enhancement" in *Bioethics*, 30, 9: 714-720.

Sparrow, Robert, 2014, "Better Living Through Chemistry? A Reply to Savulescu and Persson on 'Moral Enhancement'" in *Journal of Applied Philosophy*, 31, 1: 23-32.

Strawson, Peter, 1962, "Freedom and Resentment" in *Proceedings of the British Academy*, 48: 187-211.

Walker, Mark, 2009, "Enhancing Genetic Virtue: A Project for Twenty-First Century Humanity?" in *Politics and the Life Sciences*, 28, 2: 27-47.

Wee, Siaw-Chui, *et al.*, 2021, "Can 'Nudging' Play a Role to Promote Pro-Environmental Behaviour?" in *Environmental Challenges*, 5: 1-13.

Wheatley, Thalia *et al.*, 2005, "Hypnotic Disgust Makes Moral Judgments More Severe", in *Psychological Science*, 16, 10: 780-784.