Why Internal Moral Enhancement Might Be Politically Better than External Moral Enhancement

By John Danaher*

Forthcoming in *Neuroethics* (special issue on the politics of moral enhancement)

Technology could be used to improve morality but it could do so in different ways. Some technologies could augment and enhance moral behaviour *externally* by using external cues and signals to push and pull us towards morally appropriate behaviours. Other technologies could enhance moral behaviour *internally* by directly altering the way in which the brain captures and processes morally salient information or initiates moral action. The question is whether there is any reason to prefer one method over the other? In this article, I argue that there is. Specifically, I argue that internal moral enhancement is likely to be preferable to external moral enhancement, when it comes to the legitimacy of political decision-making processes. In fact, I go further than this and argue that the increasingly dominant forms of external moral enhancement (algorithm-assisted enhancement) may already be posing a significant threat to political legitimacy, one that we should try to address. Consequently, research and development of internal moral enhancements should be prioritised as a political project.

Keywords: Moral Enhancement; Extended Mind; Ethical Parity Principle; Political Legitimacy;

* Lecturer in Law, NUI Galway

Introduction

Technology could be used to improve moral behaviour but it could do so in different ways. Some technologies could augment and enhance moral behaviour *externally* (so to speak). Imagine if your smartphone or smartwatch had an app that issued moral guidance for everyday moral dilemmas and that it did so in a way that maximised the likelihood of you acting on that guidance. That would be an example of an external moral enhancement: it would use external cues to push and pull you towards morally preferred actions. Other technologies could augment and enhance moral behaviour internally. Imagine being nasally dosed with oxytocin to enhance your perception and feelings of empathy,[1] or having some brain implant directly augment the capacities underlying moral reasoning. These would be examples of internal moral enhancements: they would try to directly augment the neural systems that enable you to capture and process morally salient information.

Assuming that both methods produce similar outcomes (i.e. assuming they both genuinely improve the conformity of human behaviour with moral requirements) is there any reason to favour one over the other? In this paper, I argue that there is. Specifically, I argue that internal moral enhancement is likely to be preferable to external enhancement when it comes to ensuring the legitimacy of political decision-making processes. In fact, I go further than this and argue that the currently dominant forms of external enhancement may already be posing a significant threat to the legitimacy of such processes. Consequently, research and development of internal moral enhancements should be prioritised as a political project. The argument works from a position of empirical uncertainty and is tailored to specific facts about our current technological predicament. It does not claim that internal moral enhancements are always and everywhere better than external ones. Rather, it claims that, based on existing trends and reasonable extrapolations from those trends, prioritising internal moral enhancement is likely to be a *better bet* than prioritising external moral enhancement.

I defend this argument in an indirect manner. I start, in section two, by further elaborating on the distinction between the two methods of enhancement and clarifying my central thesis. In section three, I look at the debate surrounding Neil Levy's ethical

[1] These are the kinds of enhancement discussed by Persson and Suvalescu [1, 2] in their contributions to the field. Though see the discussion in section 4 below.

parity principle. According to Levy, if we embrace something like the extended mind hypothesis, we should look on external and internal enhancements in a similar light: there is no reason to think that one method is necessarily morally superior to the other. Levy tries to use this parity-thesis to make us more receptive to internal forms of enhancement. Critics pounce on this by enumerating what they take to be important moral differences between the two forms. I side with the critics, but in an unusual way. I agree that there are likely to be important differences between the two methods of moral enhancement but these differences should cause us to favour internal forms over external forms. In section four, I build upon this argument by highlighting the possible benefits of internal methods of enhancement in the political sphere. Finally, in section five, I respond to four worries about my call to prioritise the development of internal methods of moral enhancement as a political project.

2. The Different Types of Moral Enhancement

Use of the term "enhancement" can sometimes be confusing. This is partly because it is a value-laden term[2] and partly because there are different purposes behind and different methods of enhancement.[3] To render my argument more perspicuous, I will try to clarify my use of the terminology with three comments.

First, and to avoid pre-judging the issue, I take it that references to "enhancement" technologies are intended to be understood in a narrow sense. In using the term, we do not assume that the technology is an enhancement *tout court* but, rather, that it is an enhancement in some particular, defined, domain. For example, we might refer to modafinil as an "enhancement drug" because it improves concentration, wakefulness and (perhaps) some other cognitive traits, but not because it improves our lives all things considered. There could be other side effects or consequences of such drug use that need to be factored-in when making such a general assessment. This means that even if we are fully convinced of the (narrow) enhancing effects of a drug like modafinil, there is still an interesting debate to be had about whether it should be used. In describing enhancements in this way, I appeal largely to a functionalist theory of

[2] The value-laden nature of the terminology is discussed in several sources, e.g. Parens [3] and Pacholcyzk and Harris [4].
[3] For an extended taxonomy, see Raus, K et al [5].

enhancement as opposed to a welfarist one.[4] In other words, I maintain that something is an enhancement if it improves the performance in one domain relative to some population-level or species-level norm of functioning in that domain, not simply if it improves someone's life relative to their previous state of existence.

Second, I assume that there are many different goals or "targets" one could have in mind when employing the "enhancement" label. These should all be appropriately singled out by some modifying descriptor. Thus, for example, use of the modifier "*cognitive*" is appropriate when discussing a technology that improves the ways in which we process and package information. In the present case, I am interested in *moral* enhancers, which I take to be anything that enhances human moral judgement and behaviour. This is a very broad definition and could include anything that develops morally-relevant[5] 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). The definition is deliberately broad. My ultimate concern in this article is with the effect of the enhancement on our behaviour (i.e. on our conformity with moral requirements) but I assume that there are many potential routes to that effect on behaviour. The goal is to contrast the internal and external routes. Also, when I talk about such technologies, I assume that they can *actually* enhance morally-relevant emotions, virtues, reasoning capacities and actions, and not merely that they are *intended* to do so. Thus, for the sake of the present argument, I am assuming that the technologies in question are successful, not merely aspirational. Again, this does not mean that when I refer to something as a "moral enhancer" I prejudge its overall moral acceptability; I merely assume that it is successful in some narrow domain of moral judgment or behaviour.

Finally, the goal of the present inquiry is to contrast different methods of moral enhancement. One could delineate between methods in a very fine-grained manner, but I will be coarse-grained by comparing the two general methods set out in the introduction (internal versus external). This coarse-grained distinction employs a spatial

[4] On the distinction, see Savulescu J, Sandberg A, Kahane G. [34]

[5] I adopt this modifier because it is simplistic to talk about an emotion or a reasoning capacity as being moral in and of itself. Emotions and reasoning capacities are only 'moral' in particular contexts.

metaphor and so raises the question: where exactly are these "internals" and "externals" located? The answer can be found in the following definitions:

> **Internal Moral Enhancer**: Any technology that enhances moral judgment or behaviour by *directly* augmenting *neural systems and processes* underlying *morally-relevant emotions, virtues, reasoning powers* and *action planning*.

> **External Moral Enhancer**: Any technology that enhances moral judgment or behaviour by changing some external factor that causally influences such judgments and behaviours.

In other words, the "internal" to which I refer is "internal to the brain"; and the "external" is anything outside of that. To give an example, a drug or brain implant that directly enhanced our consciously represented feelings of empathy towards a victim of violence, or which dampened activity in neural circuits associated with biased or prejudicial reasoning (and thereby enabled us to adopt a less biased and more impartial perspective) when deciding how to distribute food aid to a group of starving villagers might[6] count as an internal moral enhancer. Contrariwise, a smartphone app that gamified charitable giving in order to nudge[7] us toward increased donation, or a piece of wearable technology that gave you an electric shock every time you were about to do something morally inappropriate might count as an external moral enhancer. This isn't a purely hypothetical example. There is a device in existence called the Pavlok that tries to improve behavior through electric shocks.[8] Similarly apps that allow us to select, but then outsource or automate certain moral performances could count as external moral enhancers. A simple example of this might be an automated communication device that sent affectionate messages to your loved ones to let them know you are thinking about them and that you love them. I take it that such communications are a common and morally virtuous performance within a relationship. There are a number of apps that facilitate this form of automated communication. They allow you to preselect messages

[6] I use these merely as examples. You may argue that impartiality or enhanced empathy would not improve moral judgment or behaviour. That is fine, simply substitute an example that does involve the enhancement of some aspect of conscious decision-making that would count as a "moral" enhancement.
[7] I use this term on several occasions throughout this article as a nod to the 'nudge' philosophy first defended by Sunstein and Thaler [37]. I refer to it throughout as I view the nudge philosophy as one of the leading forms of external moral engineering.
[8] There are already devices of this sort. An example would be the Pavlok, which is designed to facilitate habit change and which is named after Ivan Pavlov, the Russian behavioural psychologist. One could easily imagine this device being used to enhance moral virtues. See http://pavlok.com/

to send at preferred intervals and some include more fully-automated machine-learning elements that try to craft messages for you based on your past choices. In those cases you make the choice to use the device and it does the moral work on your behalf.[9]

The examples just given of external enhancement are limited in an important way. One could argue that external enhancers (moral or otherwise) are ubiquitous insofar as virtually every form of human technology or cultural innovation tries to augment and enhance our ability to interact with the world. This ranges from traditional methods of education, cultural ritual and belief, to communications technologies and artificial intelligence. Those that do so successfully could all count as external enhancers. This point has been made by many participants to the enhancement debate.[10] In this particular article, however, I am taking a deliberately narrower view of external moral enhancement. I conceive of it as being underscored by contemporary information communications technology, specifically the combination of smart devices and artificially-intelligent algorithms. That's why all the examples just given were of this form.

I justify this narrower focus on the grounds that such technology is now nearly ubiquitous and deeply integrated into everyday decision-making processes and it has effectively become the dominant medium for external behavioural modification. When I wish to catch a train, plan a journey, order my shopping, organise my calendar, talk to my friends, plan my workouts and manage my finances, I rely on the information fed to me by an array of algorithms, and on the communications possibilities made available to me by the internet. When I want change my behaviour by adopting an exercise plan or alter my diet, I will often turn to one of the panoply of apps that are dedicated to helping me do this by tracking my behavior, encouraging me to share it with peers, and prompting me to do various things in an effort to become a healthier and more virtuous person. Likewise, it is increasingly true that when governments want to find potential terrorists or catch tax cheats, they now turn to big data analytical systems to help them sort through the information and highlight useful patterns within. The deep integration of this type of technology into our lives justifies the narrower focus. This means that the argument I make is tailored to the differences between these ICT-based systems of external enhancement and the internal, neurochemical or neurotechnological

[9] Examples of such apps include: BroApp; Romantimatic; and Google's Allo. These examples are discussed by Selinger [35, 36]
[10] Harris [6] and Agar [7]

equivalents. It is not really about external enhancement more generally. All that said, I will occasionally refer to a broader conceptualisation of external enhancement, particularly when considering criticisms of my position.

With these terminological clarifications out of the way, I can state my central thesis with greater precision. My claim is that internal moral enhancers are likely to be morally and politically preferable to external moral enhancers. Indeed, I believe that the excessive use of external moral enhancers may be posing a threat to the legitimacy of political decision-making processes, and that this threat may be such that the development of an appropriate and counterbalancing set of internal moral enhancers ought to be prioritised. This argument will rests on two claims about the likely effects of internal enhancement on human moral reasoning. The first is that internal enhancement is less likely to bypass explicit (consciously represented) moral reasoning processes (which are crucial to the legitimacy of political decision-making). The second is that even if internal enhancement does bypass such systems, internal *automaticity* is more valuable than external *automation*. Because increased use of external enhancement is likely to require significant amounts of automation, it is less valuable than its internal equivalent. This is an empirical argument, made with explicit recognition of the uncertainty of future technological development. The argument is not that internal enhancement will definitely be superior in political terms, but rather that its prioritization is likely to be a good bet.

I will defend this argument in an indirect manner by first considering a potential problem with its framing. In its summary formulation, the argument seems to assume that there is an important distinction between what happens inside the body and what happens outside it. The problem with this thesis is that it appears to butt up against one of Neil Levy's central claims about neuroethics, namely his "Ethical Parity Principle". So, as a first step towards defending my thesis, I will consider this principle and its discontents.

3. The Ethical Parity Principle and its Discontents

Levy's Ethical Parity Principle (EPP)[11] tries to use the ubiquity of external enhancers (moral or otherwise) in making the case for internal enhancers. In doing so, it enters some controversial metaphysical waters by appealing to Clark and Chalmers's Extended Mind Hypothesis (EMH).[12]

The EMH claims that the human mind isn't all in the head. Among mind-brain physicalists,[13] it is generally accepted that the brain plays a central role in instantiating the human mind. Which parts of the brain are most important, and how exactly they go about instantiating the human mind is a matter of considerable debate, but that parts of the brain are central to the story is not. The EMH offers a mild corrective to this point of view. In doing so, it derives support from the functionalist theory of mind.[14] According to this theory, mental states and processes such as *remembering*, *thinking*, *believing* and *intending* are best understood in abstract mechanistic terms.[15] They should be viewed as roles or locations within a mechanism performing some function or set of functions. To put it another way, mental states and processes are not to be understood in terms of the particular physical processes, elements or outputs in which they seem to be found, but in terms of their abstract causal relations with sensory inputs, other mental states and processes, and behavioural outputs.[16] One of the apparent consequences of this functionalist view is that mental states and processes are *multiply realisable*. Any mechanism, provided it instantiates the right set of abstract causal relationships, could count as a mind, with digital computers often being singled out as the best possible alternative medium.

The EMH draws out some further implications of the functionalist view. Clark and Chalmers (among others)[17] argue that if mental states are *multiply realisable* then it is also probably the case that they are *jointly-realisable*. In other words, if a mind could be instantiated by a human brain or an appropriately organised technological artifact like a digital computer, it is also possible that it could be instantiated by the combination of a human brain and a technological artifact. This is illustrated by Clark

[11] See Levy [8, 9, 10] for a discussion of the parity principle
[12] Clark and Chalmers [11] and Clark [12]. I use "hypothesis" where some refer to "thesis". I don't believe the distinction counts for much since no one denies that claims about mental extension are, to a considerable degree, metaphysical and not easily amenable to empirical refutation.
[13] And, indeed, many dualists.
[14] The EMH is certainly most at home with the functionalist theory, but some have recently argued that a certain version of the EMH is compatible with nearly every theory of mind. See Farkas [14]
[15] I use the term mechanistic here in light of the abstract characterisation of mechanistic explanations found in work by philosophers of science like Craver [15].
[16] Jaworski [16] p. 136-140
[17] Levy himself notes that others have defended similar views, see Levy [9]

and Chalmers's famous Otto-Inga thought experiment.[18] The thought experiment compares a man named Otto who has a memory impairment and uses a notebook to store information that he uses to remember where he needs to go, with a woman named Inga who has no memory impairment and simply uses an appropriately activated part of her brain. Clark and Chalmers argue that both individuals have roughly equivalent mental processes of remembering, it just happens to be the case that Inga's memory is all in her head, whereas Otto's is spread out between his brain and some particular technological artifact (the notebook).

There is more to it than that, of course. Clark and Chalmers add a number of conditions that need to be satisfied before a technological artifact will count as part of one's extended mind. For instance, they claim that the artifact must be readily *accessible* and its contents, more or less, *automatically endorsed*. Critics dispute the claim that the mind can extend into such artifacts, or that these conditions are sufficient for something to count as part of the mind. I propose to sidestep these larger metaphysical questions in this article. For the sake of argument, I will accept that the EMH could be true — that the mind can extend into external technological artifacts. What matters is whether the extension of the mind into technological artifacts is ethically significant.

This is what Levy tries to assess. He suggests that the extended mind hypothesis implies that internal enhancement is on a par with external enhancement. The claim must be understood in context. Levy draws attention to the EMH because of the persistent belief among critics of enhancement that there is some in-principle objection to the use of technologies that directly augment or enhance parts of the human brain. One such critic is Erik Parens who has argued that "means matter morally" when it comes to assessing the probity of enhancement. In particular, he has suggested that external forms of enhancement are less ethically objectionable than internal forms.[19] Levy's counterpoint is that, if the EMH is true, such in-principle objections are misguided. For, if the EMH is true, then we are always enhancing the human mind through the use of technology. What matters is not how we go about doing it but, rather, which precise forms of enhancement we should promote or discourage. As Levy himself puts it:

---

[18] Clark and Chalmers [11].
[19] Parens still expresses allegiance to the "means matter morally" point of view, though he has moderated his position somewhat. See Parens [17], particularly chapter 4.

> "*Much of the heat and the hype surrounding neuroscientific technologies stems from the perception that they offer (or threaten) opportunities genuinely unprecedented in human experience. But if the mind is not confined within the skull…[then] intervening in the mind is ubiquitous. It becomes difficult to defend the idea that there is a difference in principle between interventions which work by altering a person's environment and that work directly on her brain, insofar as the effect on cognition is the same; the mere fact that an intervention targets the brain directly no longer seems relevant."*
>
> (Levy [10], 291)

This is what leads him to formulate the Ethical Parity Principle (EPP). That principle comes in two forms,[20] strong and weak. The strong form claims that, if the EMH is true (as Levy believes it to be), then "alterations of external props used for thinking are (*ceteris paribus*) ethically on a par with alterations of the brain".[21] The weaker form does not rely on the actual truth of the EMH, but merely on the fact that the mind is deeply embedded in a network of external causal factors. It holds that, because of this embedding, "alterations of external props are (*ceteris paribus*) ethically on a par with alterations of the brain to the precise extent that *reasons* for finding alterations of the brain problematic are transferable to alterations of the environment in which it is embedded."[22]

It is important to note the two different uses to which Levy puts his parity principles. The first (evinced by the strong EPP) is to soften people up to be more accepting of enhancement technologies that directly augment the human brain. Since we already accept technologies that indirectly augment the brain (via changes in the external environment) we should be more willing accept technologies that do so directly. The second use to which Levy puts the parity principle (evinced by the weak EPP) is to heighten people's awareness of the ethical problems associated with manipulating the external environment. If we would object to deleting someone's

[20] DeMarco and Ford [18] have recently offered some modifications to the weak form and reasons to reject the strong form. I return to these critiques below.
[21] Levy [8] as quoted in DeMarco and Ford [18]
[22] The quote comes from DeMarco and Ford [18]

memories through lobotomy then we should, perhaps, feel the same about deleting the information stored on someone's smartphone.

As it happens, I am sympathetic to both of these uses. But I am also sympathetic to the views of Levy's critics. I believe, along with them, that there are likely to be important moral differences between internal enhancements and external enhancements. These are not 'in-principle' differences; they are 'in fact' (or likely in fact) differences. But unlike Levy's critics, I believe that these moral differences are such as to render internal enhancement preferable to external enhancements. In saying this, I concur with Levy's weak parity principle and think that the reasons for objecting to internal methods can transfer over to the reasons for objecting to external methods and vice versa. Indeed, this parity of reasoning is critical to the argument I wish to make. My claim is that the reasons for favouring internal moral enhancement are precisely the same reasons for disfavouring the external methods. In other words, the internal methods are likely to supply something that external methods are likely to take away.

To start this argument, let's consider some of the ways in which internal and external cognitive systems are different. To illustrate it makes sense to work with the example of internal and external memory systems since this has been widely discussed in the literature on the EMH.[23] Internal memory uses brain-based systems to store and activate memory recall. External memory systems use some combination of the brain and an external storage device like a notebook or a digital personal assistant. Although both can be used to the same ultimate effect (recollection), there do appear to be morally salient differences between the internal and external systems. Three of these differences have been identified by critics of the EMH and EPP. They are:

> **Dynamic integration**: Internal memory is a dynamic, not a static, phenomenon. The information stored in Otto's notebook or on a smartphone is static.[24] Once inputted, the information remains the same, unless it is deliberately altered. Internal memory is not like this. As is now well-known, the brain does not store information like, say, a hard disk stores information. Memories are dynamic. They are changed by the act of remembering. They also integrate with other

[23] See - DeMarco and Ford [18]; and Michaelian, K [19]

[24] Some people dispute the truth of this, at least when it comes to more modern smartphone apps which can integrate informational changes in a more systematic manner but even in those cases the dynamism of the device is not directly integrated into the subject's cognitive framework. See Farkas [14] for a discussion.

aspects of our cognitive frameworks. They affect how we perceive information, and how we desire and plan action. External props do not have phenomenologically similar effects. They may eventually change how we think and view the world, but these effects are more attenuated. Internal memory is more closely coupled to these other phenomena. Consequently, tinkering with internal memory could have a much more widespread effect on how we understand and interact with the world than tinkering with external memory. This carries with it a greater risk of harm or loss (and potential benefit), which is presumably ethically significant. This 'dynamic integration' difference is likely to be true for many, if not all, internal cognitive systems. We know that the brain is a massively interlinked network of neurons and glial cells; it stands to reasons that tinkering with particular components of the network will have knock-on effects elsewhere in the network.

Fungibility: External memory props may be more easily replaceable than internal memory. If I destroy your smartphone or your notebook, you can always get another one. And although you may have lost some of your externally stored memories (maybe some pictures and messages) you will still be able to form new ones. If, on the other hand, I destroy your hippocampus (part of the brain network needed to form long-term memories), I can permanently impair your capacity to acquire new long-term memories.[25] Again, this difference in fungibility seems like it is ethically significant. By tampering with internal systems I can add or take away something of serious long-term significance. By tampering with external systems this risk (or benefit) is lessened.

Consciousness: Another obvious difference between internal and external memory is the degree to which they are implicated and represented in conscious experience. Consciousness is usually deemed to be an ethically salient property. Entities that are capable of conscious experience are capable of suffering and hence capable of being morally wronged. What's more, the nature and quality of one's conscious experiences is often thought to be central to living a good life. Although the information stored or function performed by an external prop may, eventually, figure in one's conscious experiences, its figuring is very different

[25] This isn't a hypothetical example either. This has really happened to some people. The most famous being patient Henry Molaison, who had part of his hippocampus removed during surgery for epilepsy in the 1950s, and was never able to form another long-term semantic memory.

from information that is stored or functions that are performed by internal systems. As noted above, internal memory can get deeply integrated into our mental models of the world, affecting how we perceive and act in that world. So if we alter an internal memory system, it could have a much deeper impact on the quality of our conscious experience. The effect could manifest in our conscious understanding of what we are doing and what is happening to us. If you take away the internal memory system, it can have a radical impact on how you understand yourself and yourself experiences.[26] Taking away an external memory store has less radical effects on conscious understanding. Again all of this looks to be ethically significant.

I think these differences are ethically significant and I agree with critics that this undermines the overall credibility of the EPP.[27] But I disagree with the subsequent tenor of the critics. Each of the three differences listed above is described in such a way as to make one more sceptical and cautious about the moral propriety of directly enhancing the brain. Each of them suggests that brain-based processes are more fragile and more precious than external processes and hence we should be more reluctant to intervene in them. To some extent this must be true (a point to which I return in section 5), but I believe that the preciousness of the brain-based processes cuts both ways. It could just as easily be the case that the preciousness of the internal systems implies that internal enhancement is much more valuable than external. In this manner, I believe that the morally salient differences highlighted by Levy's critics can actually be used to support his two primary goals, namely to reduce opposition to direct neuroenhancement and to heighten our appreciation of the ethical problems associated with external enhancements. It is time to defend this view.

4. Why Internal Methods are Better

The essence of the view is as follows: internal methods of moral enhancement are likely to be preferable to external methods, precisely because internal methods are more likely to be directly integrated into how we consciously perceive and understand the

[26] Again this would seem to be obviously true for famous amnesiac patients like Henry Molaison. Reading case reports on his life post-surgery suggests he had a radically different understanding and awareness of his experiences post-surgery than pre-surgery. He lived in a perpetual present after the surgery.

[27] Unless it is significantly revised, as per DeMarco and Ford [18].

morally salient information in the world around us. Contrariwise, external methods are likely to be less favourable because they are more likely to bypass conscious perception and understanding. This is because external methods are likely to present us with moral action-recommendations and nudge us to follow those recommendations through rewards for good behaviour. They are consequently likely to push and pull us towards morally appropriately outcomes; not to build our capacities for moral emotion, virtue and reason. This is bad from a political perspective because a central commitment of liberal democratic governance is to decision-making systems that engage with people as moral agents (as agents capable of understanding and making use of moral reasons for action) and not as mere passive recipients of the benefits of better moral outcomes. But it is the creation of such passive recipients that is likely to result from excessive reliance on external moral enhancers. I also go a step further and argue that even if internal enhancement bypasses conscious reasoning processes on some occasions, it is better to have *internal automaticity* than it is to have *external automation*. And since external automation is likely to result from reliance on external enhancement, we have an additional reason to disfavor this method.

This argument will be controversial to many. Indeed, there is already a strain of thought within the moral enhancement debate which claims that internal forms of enhancement are more likely to bypass conscious moral reasoning capacities.[28] It is exactly this view that I call into question.[29] I do so for two main reasons. First, I think that the differences between internal and external mental systems – highlighted above in the discussion of Levy's EPP – suggest that internal enhancements are more likely to have effects that are directly integrated into the ways in which we perceive and understand the world. Second, I also believe that the dominant, ICT-based modes of external moral enhancement are far more likely than the internal methods to bypass our conscious moral reasoning. This is bad news, politically speaking.

A brief excursion into political theory is needed to defend this argument. One of the central concepts in contemporary political theory is that of *legitimacy*.[30] The fundamental assumption in liberal democratic states is that all individuals are moral equals. This means that no one individual possesses coercive moral authority over another as a matter of natural right. This assumption creates a problem insofar as some

[28] See Schermer and Focquaert [43]
[29] I'm not novel in this observation. Maslen, Pugh and Savulescu [44] point out that Schermer and Focquaert ignore ways in which internal enhancement doesn't bypass conscious reasoning.
[30] See Peter [20]

degree of coercive authority is necessary for a mutually advantageous society to exist.[31] Indeed, political institutions largely get by on their ability to make decisions with some coercive effect. The challenge is to figure out how can they do so without compromising the fundamental moral assumption of a liberal society.

Legitimacy is the answer to this question. Legitimacy is the property that coercive public decision-making processes possess when they are morally entitled to exercise coercive authority. Some conceive of legitimacy as an 'all-or-none' property. That is to say, decision-making processes are either legitimate or they are not. This is a position I reject. I believe that there can be greater or lesser degrees of legitimacy. This does not rule out the idea that there is some minimum threshold of legitimacy that all public decision-making processes must cross. It merely adds to this the idea that decision-making processes that cross this threshold can be better or worse at respecting the moral equality of their subjects. The assumption underlying this 'degree-based' view is that morally better political communities are characterised by a greater effort to ensure the greater legitimacy of their decision-making processes.

There are several different accounts of what exactly it is that imbues such processes with legitimacy. Broadly speaking, there are three schools of thought. First, there is the *pure instrumentalist* school of thought, which argues that a procedure gains legitimacy solely in virtue of its consequences: procedures are instruments that have normative aims (reducing crime, increasing well-being etc.), and the better they are at achieving those aims, the more legitimate they are. Second, there is the *pure proceduralist* school of thought, which thinks that it is difficult to know what the ideal outcome is in advance and hence there is a need for our procedures to exhibit certain outcome-independent virtues (Peter 2008). For example, pure proceduralists might argue that our procedures should create ideal speech situations, allowing for those who are affected by them to comprehend what is going on, and to contribute to the decision-making process (Habermas 1990). Finally, there is the *mixed* or *pluralist* approach to legitimacy, which thinks that some combination of procedural and instrumental virtues is needed. This is the approach I tend to favour because I agree that respecting those affected by procedures, by ensuring meaningful participation, is politically virtuous. I also agree with proceduralists that it is often difficult to identify preferred outcomes in

[31] This is the classic Hobbesian view, dressed up in the language of contemporary analytic political philosophy. See Gaus [25] for an extended discussion.

advance, hence we have to trust a particular method or procedure. At the same time, I think we cannot completely ignore outcomes. A wholly virtuous procedure that resulted in death and immiseration for every citizen would surely be less legitimate than one that resulted in greater happiness and wellbeing. Acceptance of this pluralist vision isn't essential for the remainder of the argument — acceptance of the need for some procedural virtues is — but I will assume the pluralist view going forward. The important point about the pluralist view is that it envisions potential tradeoffs being made between instrumentalist and proceduralist virtues.

If we accept this view, something interesting happens when we turn our attention to the moral enhancement debate. Political decision-making processes often have explicit moral aims and purposes. For example, when passing legislation dealing with things like environmental protection, criminal behaviour, and personal rights and freedoms, legislative bodies are clearly tasked with analysing and evaluating a number of sensitive and complex issues of public morality. If we took a purely instrumentalist approach to the legitimation of such decision-making, we would presumably be happy if the legislative body reached the best possible moral outcome in relation to those issues.[32] It would not matter exactly how this outcome was reached. It wouldn't matter whether the methods bypassed conscious moral reasoning or not. For instance, we could happily use external devices such as an electric-shock administering bracelet to ensure that individual legislators voted the (morally) right way on a particular piece of legislation. That would *enhance* the overall legitimacy. External moral enhancement would be just as good as internal moral enhancement from the instrumentalist perspective. But as soon as we start factoring in proceduralist considerations, the analysis changes. Typically, proceduralist legitimacy conditions include things like transparency, participativeness and comprehensibility. It is said to be important not merely that the legislative (or other public) body produces the morally best outcome, but that it do so in way that allows for both individual legislators and the broader public to participate in and understand the decision-making process. This means that they can appreciate the evidence used to guide the decision-making; that they can appreciate the connection between that reasoning of the policy outcome; and, in appropriate cases, that they can actually contribute to the decision-making procedure by offering their own

[32] 'Outcome' must be understood broadly here so as to include outcomes in the traditional utilitarian sense (i.e. what happens to the individuals affected) and the mix of default rules or principles that might be selected by the decision-making authority (more akin to rule utilitarianism).

evidence/testimony and presenting their own arguments.[33] In other words, it is important that they be allowed to add to the set of moral reasons deemed relevant to the decision, and that they be able to weigh and assess the moral considerations for themselves.

It is my contention that when it comes to ensuring meaningful participation, external enhancers would no longer be sufficient to improve the overall legitimacy of the legislative procedure. Internal enhancers are likely to be needed for that. The problem is that the dominant form of external moral enhancer is, as I suggested earlier, the moral enhancer that is based on information communications technology (ICT). These enhancers typically take the form of behaviour-tracking and behaviour-recommendation apps or devices that 'ping' you with suggested actions, and reward you for following their lead. These apps and devices are now characterised by their reliance on mass surveillance, data tracking, data mining, and machine learning to deliver valuable information to their human users. My claim is that these technologies are not conducive to legitimacy-enhancing human understanding and participation.[34] Thanks to the growing network of data sensors and collection devices, we are creating vast datasets that are beyond the comprehension and understanding of individual human beings. Algorithmic assistance is needed to make sense of these datasets. The problem is that contemporary data-mining and analytical algorithms are not easily transparent to human understanding. They are created by teams of engineers; they rely on machine learning methods that are not readily 'interpretable' by those engineers; and they are grafted into complex ecosystems of other algorithms that are exceptionally difficult to reverse engineer.[35] This can result in useful recommendations – maybe even ones that ensure greater conformity with moral norms – but it is corrosive of political legitimacy. It is a world of 'black-boxed' behaviour modification. We know what we are being told to do, but we don't know why.[36]

And that's not the only problem. Even if the devices are transparent and understandable by human users, they often rely for their utility on well-known psychological biases and heuristics. Where a device is used to outsource the hard labour of moral reasoning and judgment, people are likely to grow complacent, tending to trust

[33] On the question of what is required for true participativeness, see Machin [x]
[34] For a lengthy defence of this view see Danaher [32]; also Burrell [41]
[35] On these problems, see Kitchin [42], Burrell [41] and Danaher [32]
[36] For a discussion of this black-boxing effect, see Pasquale [24], Burrell [41], and Danaher [32]

the judgment of the technology over their own. They are also likely to be prone to degeneration effects.[37] The more they use the devices, the less capable they become. Thus external moral enhancers are not likely to increase their capacity for moral wisdom; they are likely to slowly erode their moral capacity.

Internal enhancers are different. Enhancers that directly augment brain-based processes will – if earlier claims about the distinction between internal and external mental systems are correct – tend to enable greater cognitive sensitivity to morally salient features of the environment around us, and greater facility in processing and acting upon that information. So for example, enhancing the morally-relevant emotions of legislators (e.g. through the use of some hormone or drug) during an important debate about social justice, should enable them to appreciate moral issues raised by this legislation that they did not previously appreciate at a conscious level. This appreciation could become integrated into their neural networks and transferred to long-term memory. It would then become part of how they perceive and understand the world. The effect is a contextual one: the internal enhancement increases capacity in a particular environmental setting. But this is qualitatively very different from an external reward or recommendation provided by a smart device. This has no synergistic contextual effect. Similarly, directly enhancing certain cognitive reasoning capacities, might allow a jury member to concentrate more on the morally salient information presented at trial and thereby improve their ability to participate in and comprehend the trial process. The enhancement here would go beyond merely improving the outcomes of these public decision-making processes. It would now include improvements to the procedural virtues of those processes. The participants would be better able to meaningfully participate in the coercive decision-making process.

That, at any rate, is the argument I wish to make. Some may not be convinced. They might argue that external moral enhancements could do an equally good job in enhancing the participativeness and comprehensibility of public decision-making processes. For instance, a smartphone app that alerted me every time the legislature was debating some morally significant legislation, and highlighted the kinds of moral argument being made in the legislative process, would surely help improve the transparency and participativeness of our public decision-making? Certainly no less so than an internal moral enhancer.

[37] For a discussion of these effects see Parasuraman R, Manzey DH [40]

But what I want to suggest is that the smartphone app example is, in actual fact, a perfect example of the type of external moral enhancer that is corrosive of legitimacy in public decision-making. Such an app is likely to rely on algorithm-based artificially intelligent systems to flag important events and highlight the morally salient considerations. It will consequently use algorithm-mediated push and pull systems of rewards and recommendations to enhance our moral awareness. This might seem like a boon at first – our attention is drawn to something we would otherwise have ignored or discounted – but over time the corrosive effects are likely to multiply. The programming code and rationales underlying this technology will be comprehensible to only an elite minority of professional programmers and computer scientists (if at all). And the system will play upon the cognitive biases alluded to above. If the judgments of the technology seem accurate, or if their effects are too difficult to ascertain, people will lapse into a state of complacency. They will only focus on what the machine tells them to focus on; they will start to accept the judgments of the technology without question. This is already happening with satellite navigation and autopilot technologies. I submit it is likely to happen with equivalent 'moral' navigation technologies. And if this occurs over extended periods of time, it may lead to the degeneration of the cognitive capacities required to make such decisions for oneself.[38] This is anything but conducive to enhanced transparency and participativeness.

Technology theorist Evgeny Morozov captures the essence of the problem by commenting on how such devices appear to reduce us to moral patients and rob us of our agency (where he refers to decisions that are "stupid, unhealthy or unsound" we can substitute "morally unsound"):

> *Thanks to smartphones or Google Glass, we can now be pinged whenever we are about to do something stupid, unhealthy or unsound. We wouldn't necessarily need to know why the action would be wrong: the system's algorithms do the moral calculus on their own. Citizens take on the role of information machines that feed the techno-bureaucratic complex with our data. And why wouldn't we, if we are promised slimmer waistlines, cleaner air, or longer (and safer) lives in return?*

[38] Carr [28] for a general defence of this view. Carr relies work by Van Nimwegen [29] on the 'degeneration effect'. This seems work highlights how the reliance on computer-aids to certain cognitive tasks can weaken one's long-term ability to engage in that task.

(Morozov [26])

A critic might interject at this point and ask: what if the internal enhancers also bypass conscious moral reasoning? What if, say, a large dose of oxytocin (or whatever) reduces legislators to drooling empathy-loving automatons? As a result they just automatically favour the morally superior outcome without any true insight into why they favour it. Isn't this just as corrosive of political legitimacy and isn't it just as likely to happen?[39]

There are two potential flaws with this argument. First, I suspect that internal methods of enhancement are still likely to have a more intimate and immediate effect on conscious awareness and understanding of moral decision-making, even if their most obvious effects are not immediately consciously accessible. Thus the more empathic legislator may not understand 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. This is less likely to happen with the external enhancer. As the external enhancers get 'smarter' they are likely to grow increasingly automated in nature. The human users who are being pushed and pulled towards the morally preferred outcomes, will be gradually pushed off the decision-making loop. They will be presented with 'defaults' that automatically guide them to the morally superior outcomes and will have to make an extra effort to deviate from those defaults. This is the rationale underlying the 'nudge' philosophy, which is influential in the design of many of the contemporary behavior change policies, apps and devices.[40] In this manner, I would argue that *internal automaticity* – i.e. the control of behavior by not-immediately-conscious neural networks – is more valuable than *external automation*. Internal automaticity ensures that human agents are a necessary part of the process and is likely to have positive downstream effects on conscious perception and understanding. External automation does not. To the extent that the logic of external enhancement is more inclined to external automation, we have another reason to disfavor it.[41]

[39] I would like to thank an anonymous reviewer for the evocative 'drooling legislators' objection.
[40] On the connection between the nudge philosophy and behavior change apps, see Lupton [38] and Frischmann [39]
[41] Note: this argument only works if the locus of control over the internal enhancement remains relatively close to the original human agent. I discuss this problem in the final section. I would like to thank an anonymous reviewer for drawing this problem to my attention.

This then is my argument as to why internal moral enhancement is preferable to external enhancement; and this is why there is such a danger in treating the two methods as potentially ethically equivalent. Internal moral enhancement, by directly targeting the neural mechanisms of moral reasoning, is more likely to become integrated into the way in which we perceive and understand the world, which in turn facilitates transparency and participation and consequently helps to mitigate the threat that external enhancements pose to political legitimacy. What's more, to the extent that external enhancements are becoming increasingly normalised in our individual and political lives, it is not just the case that we should prefer internal enhancement but that we should actually prioritise its development. Such a project may be critical to the survival of the liberal democratic political framework.

5. Objections and Replies

To briefly recap, I have argued that critics of Levy are right to say that internal and external moral enhancements are not ethically on a par. There are important differences between the two, particularly when we consider the role of moral comprehension and understanding in the legitimation of public decision-making. But like Levy (and unlike some of his critics) I have argued that these differences should increase our willingness to accept internal forms of enhancement. In what remains I will consider four objections to this line of reasoning.

The first objection takes further issue with my claim that external forms of enhancement pose a threat to legitimacy. One could argue that algorithm-assisted decision-making doesn't actually hinder our ability to understand and participate in public decision-making. It all depends on the modality used. If a device simply shocks you or rewards you, it may do little to deepen your moral reasoning, but it is hard to see why issuing moral recommendations is so bad. An analogy with more traditional and widely accepted forms of 'external' enhancement might help to make the point. Books are a kind of technology, and books by certain moral philosophers are sometimes filled with recommendations about how to act in a morally appropriate manner. Peter Singer, for example, is often excoriating in his criticism of western moral complacency toward those in the developing world. He recommends that we give more of our money to charity than we are currently inclined to do. I am often persuaded by his

recommendations. But surely reading his book doesn't compromise my ability to comprehend and understand moral reasoning. If anything, it augments it by providing me with a fresh perspective on moral decision problems. Couldn't a digital moral assistant operate in much the same way?

Maybe it could, but there are several things worth noting here. Books typically present us with arguments in favour of particular recommendations. Reading those arguments engages the process of conscious moral reasoning; it does not bypass it. It is part of a general process of moral education where the agent develops the skills and capacities for improved moral behaviour. If digital assistants function in this manner, then there can be no real objection to them. But my central contention is that they are unlikely to do so. Part of the seductive appeal of such devices is that they allow us to outsource much of the reasoning process, relieving us of a cognitive burden. 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. Indeed, this is often considered to be a boon by the creators of such technological enhancements.[42] Human reasoning is often sloppy, biased and error-prone. Our conventional human brains are already drowning in information, and easily distracted. Bypassing this conventional system through such moral outsourcing is part of the raison d'etre of external moral enhancements. But this is the very thing that makes them likely to pose a threat to legitimacy. This is an empirical hypothesis – and by raising the problem now perhaps this tendency can be avoided – but I believe this will be difficult to do due to the intrinsic nature of the technology and the ideological forces underlying its development. As long as those forces remain in place, the problem will persist.[43]

A second objection takes issue with my claim that this means that we ought to prioritise the development of internal moral enhancers. Surely, there are other ways to solve the problem posed by external enhancers? For example, if the threat is so severe perhaps we could institute a legal ban that forbids the use of such devices in public decision-making processes, or that, more drastically, bans them completely. This is certainly within the bounds of possibility, but there are two counterpoints worth noting. First, the argument I am making is largely an intramural one, concerned specifically

[42] See, for example, Google's Chris Urmson, director of the self-driving car project, and his desire to remove the steering wheel from the self-driving car due to the temptation for interference by error-prone humans – 'Google plans to build car with no steering wheel', *The Associated Press* 28 May 2014.
[43] This problem is discussed at slightly greater length in Danaher [32]

with the merits competing types of moral enhancement. My claim is not, necessarily, that we should be prioritising internal enhancement *all things considered*, but, rather, that if we are proceeding with moral enhancement at all, we should prioritise the internal type over the external type. That said, I do believe that the horse has already left the proverbial barn on this score. We are already actively developing and implementing external technological moral enhancements of the sort described above. Stopping their use now could prove difficult due to the enormous economic and social power of the tech industry. Furthermore, it is not clear that we should wish to entirely halt the development of external moral enhancements. The improved moral behaviours and outcomes made possible by such devices are sometimes to be welcomed. If we could secure them without compromising political legitimacy it would be all for the better. In this context, prioritising the development of internal enhancement may help us to have the best of both worlds

A third, and related, objection takes issue with my claim that we ought to prioritise the development of internal moral enhancements, only this time it does so due to the risks associated with such methods of enhancement. Indeed, the objection goes further by claiming that these risks are such that external methods ought always to be preferred. Nicholas Agar (in a slightly different context) has opined that methods of enhancement that are not directly integrated into human biology[44] are preferable to methods that are directly integrated. This is because they are safer and more effective. He makes his argument using a thought experiment.

> **The Pyramid Builders**: Suppose you are a Pharaoh building a pyramid. This takes a huge amount of back-breaking labour from ordinary human workers (or slaves). Clearly some investment in worker enhancement would be desirable. But there are two ways of going about it. You could either invest in human enhancement technologies, looking into drugs or other supplements to increase the strength, stamina and endurance of workers, maybe even creating robotic limbs that graft onto their current limbs. Or you could invest in other enhancing technologies such as machines to sculpt and haul the stone blocks needed for construction. Which investment strategy do you choose?

[44] Agar [7], chapter 3.

The question is strictly rhetorical. Agar's point is that the second method is obviously preferable to the first. It is the method that we have been using for centuries, and it doesn't face the same integration problem faced by the first method. Any enhancement that integrates directly with human biology must accept and work with the limitations of that biology. This is a tricky process since the human body is a delicate and complex system, evolved over millions of years. Tweaking or augmenting one aspect could have any number of deleterious and unanticipated side effects. Given this risk, we should probably always opt for the non-integrated methods. As Agar himself puts it:

> "*Those who seek [good outcomes][45] by internalizing enhancement face a challenge that those who externalize do not. They face a problem of integration. They want to make enhancements that are part of human bodies and brains. The enhancement must be directly integrated with existing human physiology. Externalizers of enhancement require only that the enhancements be operable by humans. They cleverly avail themselves of efficiencies enabled by biological design…The policy of externalizing enhancement may be less satisfying from the perspective of a worker who would like to brag about how strong he is. But it's likely to lead to speedier pyramid construction."*
>
> (Agar [7], 48)

How can the defender of internal enhancement respond to Agar's challenge? Two points seem relevant. First, we should acknowledge that crafting successful internal moral enhancers will be a risky business. I am certainly not trying to argue that we should pursue the development of such technologies in a reckless fashion. We should work slowly, taking the necessary precautions, gradually building and developing methods that pose a minimal risk to the human users. Nevertheless — and this is where the second point comes in — we shouldn't confuse the claim that internal enhancement is risky with the claim that external methods are preferable. The problem with Agar's objection is that he moves too quickly from the former claim to the latter. The argument I presented in the previous section gives reason for blocking that inference: internal

[45] In this section, Agar speaks specifically about enhancement in terms of access to *external* goods. This is due to the fact that his overall argument against radical forms of human enhancement is premised on a distinction between internal goods (i.e. goods associated with the intrinsic properties of particular kinds of actions) and external goods (i.e. goods solely associated with the outcomes of those actions). I have omitted this reference on the grounds that it would needless distract from the point I wish to highlight. I don't believe that this does any great violence to Agar's meaning.

methods are still risky, but they are not always inferior to external methods. If the latter pose a threat to the legitimacy of public decision-making procedures, there is less reason to prefer them to the former. Indeed, if they do pose such a threat, there could be a reason to deprioritise them and try to develop the alternative methods in a safe and responsible manner. That is all I am claiming here.

This brings us to the fourth and final objection which takes issue with the argument I presented on the grounds that it neglects other problems with moral enhancement. In particular, it neglects the standard objection that the use of internal enhancement technologies threatens to undermine moral autonomy and moral authenticity, whereas external devices can keep one's autonomy intact. If my smartphone issues me with moral recommendations, I can choose not to follow them; but if an internal moral enhancer becomes integrated into my cognitive and emotional reasoning capacities, the autonomy is taken away from me.[46] Consider, for instance, Savulescu and Persson's God Machine thought experiment. It asks us to imagine a future in which the science of morality is virtually complete. Every human being has genetically modified neurons that emit light signatures that can be picked up by a communications network. The signals are then processed by a central computer that is able to modify individual moral behaviour in an interesting way. As they describe it:

> *The Great Moral Project was completed in 2045. This involved construction of the most powerful, self-learning, self-developing bioquantum computer ever constructed called the God Machine. The God Machine would monitor the thoughts, beliefs, desires and intentions of every human being. It was capable of modifying these within nanoseconds, without the conscious recognition by any human subjects.*
>
> *The God Machine was designed to give human beings near complete freedom. It only ever intervened in human action to prevent great harm, injustice or other deeply immoral behaviour from occurring. For example, murder of innocent people no longer occurred. As soon as a person formed the intention to murder, and it became inevitable that this person would act to kill, the God Machine would intervene. The would-be murderer would 'change his mind.'*

[46] I am indebeted to Christoph Bublitz for encouraging me to clarify this objection.

(Savulescu and Persson [1], 10)

Although the authors try to defend the use of the God Machine and argue that it would not necessarily involve an affront to moral autonomy and authenticity, many are less sanguine. The notion that our moral choices could be altered, without our awareness, by a powerful centralised computer seems like a significant insult to personal autonomy.[47] Furthermore, the problem from the present perspective is that the God Machine, as described, seems to constitute an internal form of moral enhancement par excellence. Admittedly, the set-up of the machine makes it appear to work from the outside, but notice how its operations are described: Our brains have already been directly enhanced, the computer then monitors brain activity and alters how people consciously reason about moral decision-problems. The murderer is described as having "chang[ed] his mind". The God Machine doesn't simply issue recommendations or provide some system of rewards and punishments. It operates directly on the biological systems underlying moral reasoning. But if it is indeed an internal form of enhancement then it looks like a particularly offensive form as it denies us our moral autonomy and independence. We are under the perpetual control of the God Machine. Surely I cannot be defending the creation of such a device?

No; I am not defending the creation of such a device — I'm not even sure that it is a credible technological possibility — but I do admit that it poses a strong conceptual challenge to my argument. A number of responses suggest themselves. First, it is worth recalling the intramural nature of the argument I am presenting. I am not trying to say that moral enhancement is desirable all things considered but, rather, that if we are going to pursue it at all the internal forms are preferable. I am sticking with that claim here. A critic might respond that the claim is now less plausible since it is conceded that internal forms of enhancement could come at the expense of moral autonomy and authenticity. But it must be remembered that external forms also come with those expenses. They treat humans as metaphorical puppets on a string, pushing and pulling them towards preferred moral outcomes. To some extent this modality of interference can preserve autonomy since we would still have the residual capacity to reject the recommendations of the algorithm. But we should not be enthusiastic about this residual respect for autonomy. This for two reasons. If the persistent use of such

[47] One can draw analogies between how the machine works and Frankfurt-style cases in the literature on moral responsibility to argue that it still facilitates moral agency.

devices has a degenerative effect on our ability to reason through moral problems for ourselves – as suggested earlier – then our capacity to exercise our residual autonomy would be necessarily weakened. Second, if there is a general pressure to remove error-prone humans from the decision making loop, and hence rely on external automating devices, they will eventually take away our autonomy and independence. What's more, they will do all this without any regard for the individual comprehension or understanding of the decisions being made.

Despite this, I do agree that if we pursue internal methods of moral enhancement, we should do so in ways that try to respect individual moral autonomy and authenticity. This will be a tricky business since the concept of autonomy is so deeply contested. Do you undermine someone's moral autonomy if you adopt genetic methods of enhancement that clearly alter their conscious moral reasoning later in their lives? I tend to think not, but others see this as a form of manipulation that undermines moral freedom.[48] Nevertheless, I also tend to think that the *locus of control* is the all-important variable here. A *locus of control* that is largely internal to the biological agent (or reasonably close to that agent, e.g. a button or switch that they can flip) seems like it would retain a sufficient degree of autonomy and authenticity. The problem with Persson and Savulescu's God Machine is that the locus of control is too distant and centralised to protect moral autonomy. It is a single moral 'god' sitting at the centre of civilisation that manipulates, controls and governs human behavior. I would argue that we should disfavor such a centralised system. The methods whose creation we prioritise should be more localised to the individual.

Conclusion

In summary, if we are to pursue a project of moral enhancement, we must be careful about the method we pick. We should be sceptical of claims that external methods are ethically equivalent internal methods. There are important moral differences between the two, particularly when it comes to the political sphere. The legitimacy of public decision-making processes is something we should seek to protect, uphold and augment. Moral enhancement might help us to do this. But if we continue to pursue and prioritise external methods over internal methods we actually risk undermining that legitimacy. This is because dominant external methods tend to bypass

[48] Pereboom [31]

or obscure the forms of conscious moral reasoning that are key to the proceduralist virtues of public decision-making. Internal methods are more likely to enhance those conscious reasoning capacities. The general political community should be invested in the project of protecting those proceduralist virtues. This is why we ought to prefer (and perhaps prioritise) the creation of safe methods of internal moral enhancement.

Acknowledgments: The author would like to thank Christoph Bublitz, Norbert Paulo and an anonymous reviewer for feedback on previous drafts of this paper.

Funding: The research for this paper was kindly funded by the Irish Research Council, New Horizons Grant.

References

1. Savulescu, J and Persson, I (2012). Moral Enhancement, Freedom and the God Machine. *The Monist* 95(3): 399-421 (page references are to the online version, available open access at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431130/pdf/ukmss-49380.pdf - accessed 3/4/15)

2. Persson, I and Savulescu, J (2012). *Unfit for the Future*. Oxford: OUP.

3. Parens, E. (2014). *Shaping Our Selves*. Oxford: OUP

4. Pacholczyk, A. and Harris, J. (2012) Dignity and Enhancement. In Nathan J Palpant and Stephen C Dilley (eds), *Human Dignity in Bioethics*. London: Routledge.

5. Raus, K et al (2014). On Defining Moral Enhancement: A Clarificatory Taxonomy. *Neuroethics* 7: 263-273

6. Harris, J. (2007) *Enhancing Evolution: The Ethical Case for Making Better People*. Princeton NJ: Princeton University Press.

7. Agar, N. (2013). *Truly Human Enhancement*. Cambridge, MA: MIT Press.

8. Levy, N. (2007a). *Neuroethics: Challenges for the 21st Century*. Cambridge: Cambridge University Press.

9. Levy, N. (2007b). Rethinking Neuroethics in Light of the Extended Mind Thesis. *American Journal of Bioethics* 7(9): 3-11.

10. Levy, N. (2011). Neuroethics and the Extended Mind. In Sahakian, B. and Illes, J (Eds) *Oxford Handbook of Neuroethics*. Oxford: OUP.

11. Clark, A. and Chalmers, D. (1998). The Extended Mind. *Analysis* 58(1): 7-19.

12. Clark, A (2010. *Supersizing the Mind*. Oxford: OUP.

13. Adams, F and Aizawa, K. (2010). Defending the Bounds of Cognition. In Menary, R (Ed) *The Extended Mind*. Cambridge, MA: MIT Press.

14. Farkas, K (2012). Two Versions of the Extended Mind Thesis. *Philosophica* 40: 435-447

15. Craver, C. (2007). *Explaining the Brain*. Oxford: OUP

16. Jaworski, W. (2011). *Philosophy of Mind: A Comprehensive Introduction.* Oxford: Wiley-Blackwell.

17. Parens, E. (1998). Is Better Always Good? In Parens (ed) *Enhancing Human Traits*. Washington, DC: Georgetown University Press.

18. DeMarco, J and Ford, P. (2014). Neuroethics and the Ethical Parity Principle. *Neuroethics* 7: 317-325

19. Michaelian, K (2012). Is External Memory Memory? Biological Memory and the Extended Mind. *Consciousness and Cognition* 21(3): 1154-1165.

20. Peter, F. (2014) Political Legitimacy. In Edward N. Zalta (ed) *The Stanford Encyclopedia of Philosophy Spring 2014 Edition* -- available at http://plato.stanford.edu/archives/spr2014/entries/legitimacy/

21. Peter, F. (2008). Pure Epistemic Proceduralism. *Episteme* 5: 33

22. Habermas, J. (1990). Discourse Ethics: Notes on a Program of Philosophical Justification. In *Moral Consciousness and Communicative Action*. Trans. Christian Lenhart and Shierry Weber Nicholson. Cambridge, MA: MIT Press.

23. Dormehl, L (2014). *The Formula: How Algorithms Solve All our Problems and Create More*. Perigree Books.

24. Pasquale, F. (2015). *The Black Box Society*: *The Secret Algorithms and that Control Money and Information*. Harvard: Harvard University Press.

25. Gaus, G. (2010). *The Order of Public Reason*. Cambridge: Cambridge University Press.

26. Morozov, E. (2013). The Real Privacy Problem. *MIT Technology Review* (available at: http://www.technologyreview.com/featuredstory/520426/the-real-privacy-problem/ - accessed 1/3/15)

27. Zuboff, S. (2015). Big Other: Surveillance Capitalism and the Prospects of an Information Civilization. *Journal of Information Technology* 30: 75-89

28. Carr, N (2015) *The Glass Cage: Where Automation is Taking Us*

29. Van Nimwegen, C. et al (2006). The Paradox of the Assisted User: Guidance can be Counterproductive. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 917-926

30. Burgos, D. et al (2007). Game-based learning and the role of feedback. A case study. *The International Journal of Advanced Technology in Learning* 4(4): 188-193.

31. Pereboom, D (2014). *Free Will, Agency and Meaning in Life*. Oxford: OUP.

32. Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy and Technology*. DOI: 10.1007/s13347-015-0211-1

33. Danaher, J. (2016). An Evaluative Conservative Case for Biomedical Enhancement. *Journal of Medical Ethics*. DOI: 10.1136/ medethics-2015-103307

34. Savulescu J, Sandberg A, Kahane G. Well-being and enhancement. In: Savulescu J, ter Meulen R, Kahane G, eds. *Enhancing human capacities*. Oxford: Blackwell Publishing Ltd, 2011:3–18

35. Selinger, E. (2014) The Outsourced Lover. *The Atlantic* 14 February 2014 – available at http://www.theatlantic.com/technology/archive/2014/02/the-outsourced-lover/283833/ (accessed on 8/7/2016)

36. Selinger, E. (2014) Today's Apps are Turning Us Into Sociopaths. 26 February 2014 – available at http://www.wired.com/2014/02/outsourcing-humanity-apps/ (accessed on 8/7/2016)

37. Sunstein, C. and Thaler, R. (2008). *Nudge: Improving Decisions about Health, Wealth and Happiness*. Yale University Press.

38. Lupton, D. (2016). *The Quantified Self*. London: Polity Press.

39. Frischmann, B. (2014). Human Focused Turing Tests: A Framework for Judging Nudging and the Techno-Social Engineering of Human Beings. *Cardozo Legal Studies Research Paper No. 44* – available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2499760 (accessed 8/7/2016).

40. Parasuraman, R. and Manzey DH. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors* 52(3): 381-410.

41. Burrell J. (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data and Society* doi: 10.1177/2053951715622512

42. Kitchin, R. (2016) Thinking critically about and researching algorithms. *Information Communications Science* . DOI: 10.1080/1369118X.2016.1154087

43. Schermer, M. and Focquaert, F. (2015). Moral Enhancement: Do Means Matter Morally? *Neuroethics* 8(2): 139-151.

44. Maslen, H., Pugh, J. and Savulescu, J. (2015). The Ethics of Deep Brain Stimulation for the Treatment of Anorexia Nervosa. *Neuroethics* 8(3): 215-230.

45. Machin, D. (2009) The Irrelevance of Democracy to the Public Justification of Political Authority. *Res Publica*. 15:103-120.