Neurointerventions: Punishment, Mental Integrity, and Intentions

Peter Vallentyne, U. Missouri

American Journal of Bioethics Neuroscience (AJOB Neuroscience),

https://doi.org/10.1080/21507740.2018.1496185, 2018

In their interesting paper, David Birks and Alena Buyx (2018) discuss the moral permissibility of using mandatory (non-consensual) neurointerventions as an alternative to incarceration, for at least some people convicted of crimes. Neurointerventions are "interventions that exert a physical, chemical or biological effect on the brain in order to diminish the likelihood of some forms of criminal offending" (p. ?). For example, a sex offender might be given testosterone-lowering drugs. The authors argue that mandatory neurointerventions are pro tanto wrong, because they involve an intentional interference with the offender's non-disvaluable mental states. They do *not* argue that such neurointerventions are always *conclusively* (all things considered) morally wrong. They allow that sometimes there may be positive pro tanto considerations that override this pro tanto impermissibility (e.g., the presence of additional benefits, or the absence of certain costs).

In this short piece, I will merely raise some issues where I disagree with their position. I will not attempt to defend my position against theirs. I develop my own view in Vallentyne (2018).

In agreement, I think, with the authors (and almost everyone else), I hold that mandatory neurointerventions are *wrongful* interferences with the person's mental states, if there is no rectificatory justification (in virtue of past wrongdoing) and no preventive justification (in virtue of possible future wrongdoing). In such cases, on my view, the intervention infringes the

individual's rights, thus wrongs her, and is therefore pro tanto impermissible. It is possible that the intervention is nonetheless conclusively (all things considered) permissible, since there may be an overriding justification for wronging the individual (e.g., it is necessary and sufficient to save eight-billion lives).

Throughout, I focus on whether a neurointervention wrongs the individual and thereby is pro tanto impermissible (as claimed by the authors). As indicated above, this is so when there is neither a rectificatory, nor a preventive, justification for the intrusion upon the individual's rights. For rectification, I hold that only victim restoration (e.g., recognition and compensation) is relevant (and not retribution or general deterrence), but I shall not address that issue. Instead, I will focus on preventive justifications, since the authors seem to focus implicitly on those. For simplicity, I focus on cases of prevention of future wrongdoing by a person who is guilty of past wrongdoing. (I do not address cases where the person has *not yet* engaged in wrongdoing or where the intervention prevents *others* from wronging).

There will be a preventive justification for mandatory neurointerventions just in case (roughly) the harms to the intervenee are *necessary for*, and *proportionate to*, the preventive benefits. The relevant benefit, I would argue, is the reduction in the expected (i.e., probability-weighted) value of the wrongful harm (of any sort) that the individual imposes in the future. For a given intervention, we must thus examine how much it reduces the expected value of her future wrongful harm and ask: (1) Necessity: Is there a feasible alternative that infringes no one else's rights and is no more costly to implement that (a) produces relevant benefits that are at least as great, but (b) is less harmful to the intervene? If so, the intervention wrongs the individual, because it is not necessary to achieve the benefit. (2) Proportionality: Is the cost imposed on the intervenee excessive relative to the benefits achieved (e.g., great suffering imposed to achieve a

trivial chance of a trivial reduction in trivial harms may be excessive). If so, the intervention wrongs the individual, because the costs to her are disproportionate relative to the benefits achieved.

Let me state some assumptions/claims of the authors, with which I agree. First, I will assume that minimally decent incarceration (with no overcrowding or risk of assault) is feasible. This is relevant, since necessity and some conceptions of proportionality are relative to what is feasible. Second, neurointerventions (at least for the foreseeable future) impose at least some costs on the intervenee (e.g., pain of an injection, bodily side effects, and mental side effects). More specifically, I also agree that some of the costs may typically involve interference with the individual's *mental integrity* (e.g., non-consensually bypassing the agent's autonomous thought in order to modify her beliefs or desires).

I originally thought (but see below) that the authors claimed that *intentionally imposed* harms are morally more significant than non-intentionally imposed harms of the same type and size. I agree that the *psychological harm* to a person is often greater when she knows/believes that the base-harm was intentionally imposed by another. I thus agree that this makes it more difficult to establish that the intervention is necessary and proportionate. The authors, however, hold something stronger: that, not only is the psychological harm typically larger in such intentional cases, but also, the pro tanto moral badness or impermissibility of the harm is stronger. There is, of course, a long tradition of holding this view (e.g., Kant), but I reject it. Moral permissibility is, I would argue, *victim-centered*, and *not agent-centered*. More specifically, moral badness and pro tanto wrongness are based on the harms to the victims, and not independently based on the harmer's mental states (what was intended, foreseen, etc.).

In correspondence, the authors pointed out that, although it is not sufficiently explicit,

their argument is meant to be *conditional* on the assumption that harms are morally more significant when they are intentionally imposed. They do not assert the assumption. So, my objection above is irrelevant.

The authors seem to hold that neurointerventions are *always* intentional interferences with mental integrity and minimally decent incarceration is not always so. The crucial question concerns the conditions under which an outcome is intentional. Of course, if the outcome is unforeseen, it is not intended (aimed at). Even a foreseen outcome need not be intended, since it may simply be a foreseen side effect (e.g., when switches the trolley track to save five lives, merely foreseeing that it will kill someone else).. An outcome is intended, roughly, just in case it is aimed at by the agent, either as end (e.g., saving the five lives) or as a means to an end (e.g., switching the track).

Following an account from William Fitzpatrick (2006), the authors claim that, if one intends X, and one knows that X is (metaphysically) *constitutive* of Y, then one intends Y. The relation of X being constitutive of Y is understood to be (1) weaker than conceptual entailment (blowing someone up is constitutive of his death, but the former does not conceptually entail the latter), and (2) stronger than causal entailment (turning the trolley causally ensures the death of five people, but it is not constitutive of their death).

The authors then argue that (for example): (1) the state of affairs of someone's testosterone being diminished by the administration of a neurointervention is *constitutive* of the state of affairs of that person being less likely to have a non-disvaluable sexual desire, whereas (2) the state of affairs of his liberty being diminished by minimally decent incarceration is not so constitutive, even if it *causally* leads to that result. Consequently, they claim, one cannot intend to administer testosterone without also intending to make the individual less likely to have a non-disvaluable

sexual desire, but one can intend to diminish her liberty by deploying minimally decent incarceration without such an intention.

Here, I must plead ignorance. I'm deeply suspicious of the moral relevance of the notion of constitution, but I don't understand it, and I'm not an expert. So, I'll merely flag this move as one that warrants further scrutiny. My inclination is to think that, if intentions matter morally, it is in the sense of psychologically aiming at a result, either as a means or as an end. Moreover, this is not (it seems to me) reducible to one's beliefs and desires (e.g., I can arbitrarily intend to hit a target, without believing that I am likely to do so and without any desire to do so). Thus, it seems to me that, whether a neurointervention involves intentionally interfering with non-disvaluable mental integrity depends critically on what the specifics of the intervener's mental states are on that given occasion. I don't understand how a general appeal to constitution can be relevant.

In sum, Birks and Buyx rightly address the moral permissibility of mandatory neurointerventions for crime prevention, and they rightly identify the relevance of the harms to the intervenee's mental integrity. I'm skeptical, however, that intentional harms are more difficult to justify than non-intentional harms, and I'm skeptical of the relevance of any metaphysical constitution relation for the determination of intentions. I have not, however, defended this skepticism. The paper is well worth a more elaborate assessment.[1]

Bibliography

David Birks and Alena Buyx (2018). "Punishing Intentions and Neurointerventions," *AJOB Neuroscience*, forthcoming.

William J. Fitzpatrick (2006). "The Intend/Foresee Distinction and the Problem of 'Closeness',"

Philosophical Studies 128: 585-617.

Peter Vallentyne (2018) "Neurointerventions, Self-Ownership, and Enforcement Rights", in *Treatment for Crime*: *Philosophical Essays on Neurointerventions in Criminal Justice*, edited by David Birks and Tom Douglas (Oxford University Press).

---

[1] I'm grateful to David Birks for his helpful comments.