AJOB Neuroscience, 9(3): 133–143, 2018

ISSN: 2150-7740 print / 2150-7759 online
DOI: 10.1080/21507740.2018.1496162

Target Article

Check for updates

Punishing Intentions and Neurointerventions

David Birks, University of Oxford
Alena Buyx, Technical University of Munich

How should we punish criminal offenders? One prima facie attractive punishment is administering a mandatory neurointervention—"interventions that exert a physical, chemical or biological effect on the brain in order to diminish the likelihood of some forms of criminal offending" (Douglas and Birks 2018, 2). While testosterone-lowering drugs have long been used in European and US jurisdictions on sex offenders, it has been suggested that advances in neuroscience raise the possibility of treating a broader range of offenders in the future. Neurointerventions could be a cheaper, and more effective method of punishment. They could also be more humane. Nevertheless, in this paper we provide an argument against the use of mandatory neurointerventions on offenders. We argue that neurointerventions inflict a significant harm on an offender that render them a morally objectionable form of punishment in a respect that incarceration is not. Namely, it constitutes an objectionable interference with the offender's mental states. However, it might be thought that incarceration also involves an equally objectionable interference with the offender's mental states. We show that even if it were the case that the offender is harmed to the same extent in the same respect, it does not follow that the harms are morally equivalent. We argue that if one holds that intended harm is more difficult to justify than harm that is unintended but merely foreseen, this means neurointerventions could be morally objectionable in a significant respect that incarceration is not.

Keywords: enhancement; harm; intention; law; morality/ethics; neuroethics

How should we punish criminal offenders? One commonly used form of punishment is incarceration. In England and Wales, the incarcerated population is over 85,000 (Berman and Dar 2016, 4); in the United States, it is more than 2.2 million (Bureau of Justice Statistics 2009). However, incarceration is a costly, and ineffective method of punishment, regardless of why we ought to punish.[1] Around two-thirds of those who receive a custodial sentence for less than a year reoffend within a year (Adebowale 2010; Durose 2014). It costs more than £40,000 a year to keep the average inmate in prison, and the cost to the UK taxpayer of reoffending is estimated to be between £9.5 and £13 billion per year. In the United States, the cost of incarceration is $39 billion per year (Adebowale 2010; Henrichson and Delaney 2012, 6).

As a result, there is growing political and scientific interest in developing an alternative method of punishment, namely, the use of mandatory neurointerventions—"interventions that exert a physical, chemical or biological effect on the brain in order to diminish the likelihood of some forms of criminal offending" (Douglas and Birks 2018, 2).[2] While testosterone-lowering drugs have long been used in European and U.S. jurisdictions as a means of preventing recidivism in sex offenders, it has been suggested that advances in neuroscience raise the possibility of treating a broader range of offenders in the future. For example, it might be possible to develop neurointerventions that increase empathy and reduce violent urges (Krakowski and Czobor 2014; Sitaram et al. 2014). Rather than solely

1. That is, incarceration is a costly and ineffective method of acting in accordance with commonly proposed reasons to punish, such as retributivist, deterrence, and rehabilitative reasons.
2. For example, the UK government proposed a program that treated persons with dangerous and severe personality disorder as psychiatric disorders. See Duggan (2011).

Address correspondence to David Birks Department of Politics and International Relations, University of Oxford, Manor Road Building, Manor Road, Oxford, OX1 3UQ, United Kingdom of Great Britain and Northern Ireland. E-mail: david.birks@politics.ox.ac.uk

using incarceration as punishment, it has been proposed that we could administer mandatory neurointerventions to an offender as a replacement for incarceration, or to reduce the length of time an offender is incarcerated (Douglas 2014; Ryberg 2012).

It is difficult to overstate the potential benefits of being able to do this. Neurointerventions could be a cheaper and more effective method of reducing recidivism. They could also be more humane. After all, if an offender is not incarcerated, and receives mandatory neurointerventions as an alternative punishment, this could enable him to maintain relationships with family and friends. The offender could continue to pursue his career, with all the financial benefits to that go with that for his dependents and to wider society through tax contributions (Henrichson and Delaney 2012, 2; Smith et al. 2017).

Nevertheless, in this article we provide an argument against the use of mandatory neurointerventions on offenders. We argue that neurointerventions inflict a significant harm on an offender that renders them a morally objectionable form of punishment in a respect that incarceration is not.[3] Specifically, we argue that administering a mandatory neurointervention constitutes an objectionable interference with the offender's mental states. This objection itself is not novel, and similar concerns have been raised elsewhere (Bublitz and Merkel 2014; Shaw 2018). However, we believe that the reasoning behind this objection has not been unpacked sufficiently and that two central aspects require further consideration. One is to emphasize the extent to which neurointerventions interfere with the offender's mental states. The interference will not only affect his mental states that relate to his offending, but will also have wide-reaching implications for his mental states unrelated to his criminal behavior. We proceed to show that this has both theoretical and practical significance for the permissibility of administering mandatory neurointerventions.

Our other contribution is to dispel a prima facie powerful response to this objection. It could be argued that incarceration also involves an equally objectionable interference with an offender's mental states. After all, it is often said that "prison changes you" (Kane 2013, 21) Moreover, there is evidence that incarcerated offenders suffer from depression and other mental health problems as a consequence of the incarceration (Haney 2013). As a result, if neurointerventions are objectionable in one respect because they interfere with a person's mental states, then the fact that incarceration also interferes with a person's mental states to the same extent would also make them morally objectionable in that same respect. According to this objection, if everything else were equal between these two forms of punishment, there would be no grounds to prefer incarceration over neurointerventions. And of course, in reality, everything else is not equal, given the aforementioned several potential benefits of neurointerventions over incarceration.[4]

We propose that even if it were the case that an offender is harmed to the same extent in the same respect by a neurointervention as by incarceration, it does not follow that the harm-doing is morally equivalent. There could be a difference in terms of intention between some of the harm-doings caused by the two forms of punishment. If one holds that intentional harm is more difficult to justify than harm that is unintended but merely foreseen, then this could account for the view that neurointerventions are morally objectionable in one respect that incarceration is not.

Before we proceed to our argument, we should make the following two clarifications. First, it is important for us to be clear that we are not arguing that administering mandatory neurointerventions on an offender is *all things considered* wrong. It might be the case that even if neurointerventions were morally objectionable in one respect that incarceration is not, it could be morally permissible (or even sometimes obligatory) to administer mandatory neurointerventions on offenders. For example, it might be thought that the harm inflicted by incarceration is so great that the harm of neurointerventions would be preferable, despite its objectionable interference with the offender's mental states. We do not take a stand on this point in the article. Rather, we have the more modest aim to show that administering mandatory neurointerventions could be wrong in one respect that incarceration is not. But it might be helpful to emphasize that even this modest aim has a significant implication, for if mandatory neurointerventions were wrong in a respect that incarceration is not, this would account for the intuition that there is something particularly troubling about the use of mandatory neurointerventions on offenders, even if it were permissible or sometimes obligatory to administer them.

Second, we should clarify that when we discuss incarceration, we are imagining it to involve a diminution of the offender's liberty, but without the overcrowding and risk of assault commonly experienced by incarcerated persons in the United States and United Kingdom (James 2013). We call this minimally decent incarceration.[5]

3. Some proponents of administering mandatory neurointerventions do not think that they should be administered as punishment, but rather, they think that they should be a rehabilitative aspect of an offender's sentence, and they view this to be distinct from punishment. For example, see Douglas (2014). Whether or not compulsory rehabilitation constitutes punishment does not concern us here. In what follows we refer to the use of mandatory neurointerventions as punishment, but our argument also applies to those who think that we ought to use mandatory neurointerventions rather than incarceration as rehabilitation.

4. Thomas Douglas makes this point (Douglas 2014, 117). See also Petersen and Kragh (2017).

5. Thomas Douglas also makes this move (Douglas 2014, 105).

TWO QUESTIONS ON THE PERMISSIBILITY OF NEUROINTERVENTIONS

The possibilities of neurointerventions have given rise to a burgeoning philosophical literature discussing their moral permissibility. In forthcoming work, a number of influential philosophers including Kasper Lippert-Rasmussen (2018), Jeff McMahan (2018), and Peter Vallentyne (2018) defend the mandatory use of neurointerventions on offenders. They address what we call the Constrained Question:

> *The Constrained Question*: Is it at least permissible to administer a mandatory neurointervention (N) on a criminal offender (C) if it has effect (X) on C, where X is the effect that C is less likely to offend?

We think that this question overlooks an important ethical implication of administering neurointerventions. Neurointerventions do not solely make the offender less likely to offend, but have other effects on the offender too. Yet almost all of the ethical debate idealizes the effects of the neurointervention to exclude these other effects. Now, to be clear, we think that this is understandable. There are legitimate philosophical reasons to isolate the effect of the neurointerventions in this sense. After all, if neurointerventions are noncontingently morally problematic, then regardless of how cheap and effective they are, they will remain morally problematic. Moreover, there are important issues surrounding the permissibility of neurointerventions even in these idealized circumstances. Indeed, much has been written in opposition to the use of these idealized neurointerventions (see, e.g., Bennett 2018; Bublitz and Merkel 2014; Shaw 2014).

Nonetheless, this idealization can only get us so far. It does not provide practical guidance for the permissibility of neurointerventions in any scenario we will face for the foreseeable future. This is because neurointerventions will not merely affect the likelihood of offending, but also they will affect other aspects of the offender's life, and these can be significant harms. This is apparent in our current practices of chemical castration. Besides the diminution of sexual desire, chemical castration has side effects such as increased body fat and an increase in male breast tissue. It also has significant effects on the offender's mental states; for example, it interferes with the offender's desire to pursue permissible sexual relationships (Briken and Kafka 2007).

It might be asked, however, why do the side effects of presently available neurointerventions matter? Even if our abilities to produce neurointerventions are presently limited in terms of efficacy and side effects, the ethical debate is focused on the potential neurointerventions of the future, and these, many take it, will not have other effects on offenders' lives.

There are good grounds to doubt this. One of the basic principles of pharmacology is that all medicines have usually unwanted side effects (Conner et al. 2012). This holds for all effective drugs, and certainly for all pharmacological treatments that have so far been developed for neurological application. Among various physical side effects, drugs targeting the brain also lead to mental, psychological, behavioral, and/or personality effects, ranging from changes in mood to altered behaviors to suicidal tendencies (Golan and Tashjian 2012).

It is true that drug development is seeking to produce drugs aimed at the brain that are so specifically targeted that they have very few side effects. However, due to the complexity of the brain and the still limited knowledge of its overall functioning, this has only been partially successful. Unless the principles of pharmacology can be wholly suspended, no effective drug targeting the brain will be possible that does not have side effects.[6] So, in order for the debate on neurointerventions to have practical import, we should instead consider a different, so far overlooked question:

> *The Expansive Question:* Is it at least permissible to administer a mandatory N on C if it has effects X, Y, and Z on C, where only effect X relates to C's offending, and effects Y and Z do not relate to C's offending?

THE HARM OF NEUROINTERVENTIONS

It might not be clear whether the Expansive Question raises any moral issues distinct from those of the Constrained Question besides the fact that it shows that neurointerventions will have a greater number of effects than generally thought. In this section, we set out how the Expansive Question raises a moral issue distinct from the Constrained Question. We think that once the extent of the harm inflicted on the offender is illuminated, it reveals a morally relevant difference between the harms inflicted by neurointerventions when compared with incarceration.

In order to show this, let's begin by setting out what we think is problematic about the effects of neurointerventions on offenders. First of all, it cannot be the fact that these effects are harmful. We accept that we ought to harm offenders by punishing them. Punishment is by definition harmful (Boonin 2008). Instead, the types of harms caused by neurointerventions are qualitatively

6. The same holds for other types of brain intervention, due to the laws of the brain's structure and physiology. Even targeted application of the smallest possible electrodes affects brain tissue that, due to the intense interconnection of the brain's functional neuronal networks, has functions other than the one intended to be changed. Indeed, deep brain stimulation, while very effective in treating, e.g., movement disorders, has for this very reason been found to have significant side effects on, among others, personality, behavior, and impulse control. Transcranial magnetic stimulation (TMS) and electrical stimulation treatments are not being used for sustained treatment, so there are no long-term data available on side effects.

different. While some effects of neurointerventions could cause suffering, there is something particularly troubling about the effects that interfere with a person's mental states, irrespective of whether they cause suffering.

Indeed, this objection has been recently made by Elizabeth Shaw (2018) and by Bublitz and Merkel (2014). On this view, neurointerventions are wrongful in virtue of a person's interest in not having at least some of his mental states intentionally altered by others in certain ways. This is sometimes called an interest in mental integrity, and it is a distinct harm to the experiential harm of suffering, and can be inflicted without any experiential harm.[7]

We harm a person's interest in mental integrity when we intentionally create or alter one of his desires through means other than engaging with that person's autonomous thought (Birks and Douglas 2017, 424–5). Some interferences with mental states do not harm a person's interest in mental integrity—for example, providing a person with arguments, or presenting evidence with the intention of altering a person's desire directly engages with the person's autonomous thought—whereas other interventions, such as hypnotism and brainwashing, create or alter desires by bypassing them.[8] We think that similarly, neurointerventions plausibly alter an offender's desire in a way that bypasses his autonomous thought. When we administer testosterone-lowering drugs to a sexual offender, for instance, in at least some cases, we do not engage with the offender's autonomous thought, and his sexual desires are altered in a way that bypass his autonomous thought.[9]

The wrongness of intentionally interfering with a person's mental states understood in this narrow sense has considerable intuitive force. There are many possible grounds for holding it. One could appeal to the value of autonomy, or the value of expressing respect, but we do not need to take a stand on this here.[10] All we need for our argument is that when A intentionally interferes with the mental states of B in way that bypasses B's autonomous thought and is contrary to B's autonomous desire, this is harmful to B in virtue of his interest in mental integrity.[11] Henceforth, when we refer to an interference with mental states, we understand it in this narrow sense.

This interest in mental integrity can explain why we might provide a negative response to the Constrained Question. We could think that it is impermissible to administer a mandatory neurointervention to an offender when it its only effect is to make him less likely to offend, as it undermines his interest in mental integrity. We might also want to draw a distinction between interfering with disvaluable and non-disvaluable mental states.[12] For instance, it might be thought that the value of a desire depends on the value of the object of the desire.[13] As a consequence, we might think that an interference with an offender's mental states is only *conditionally* objectionable. It is one thing to interfere with a person's disvaluable mental states, such as the desire to have sex with children. It is quite another thing to interfere with a person's valuable mental states, such as the desire to have sex with a consenting adult.[14] Indeed, it is possible to hold that it is only the latter interference that is morally objectionable, and not the former.[15] Or at least, one could hold the weaker view that the latter

7. Thomas Douglas notes that this is the most promising objection to the use of mandatory neurointerventions (Douglas 2014, 119).

8. We could explain this by holding an historical account of autonomy. See for instance, Christman (1991) and Ekstrom (2005). Broadly, on these accounts, a desire having the correct origins is a necessary and sufficient condition for it to be autonomous. A desire that is created through argument would have the correct origin to be autonomous, while a desire created through brainwashing would not. This also tracks Neil Levy's distinction between direct and indirect brain manipulation. See Levy (2007, 70).

9. Testosterone-lowering drugs would not harm a person's interest in mental integrity in cases where the offender autonomously wants to have his sexual desires altered.

10. For the former, see Bublitz and Merkel (2014). For the latter, see Shaw (2018). We think that a plausible basis for this interest is connected to Oshana's relational view of autonomy (Oshana 2006).

11. It is also possible that this interest has sufficient weight to hold others to be under a duty, and thus on the interest theory, we can say that that this interest is protected by a *right*. We are not committed to holding a claim as strong as this, but it should be noted that such a right has been defended elsewhere (Bublitz and Merkel 2014).

12. To clarify, henceforth when we refer to a valuable or disvaluable mental state, we mean *all things considered* valuable or disvaluable. When a desire is *all things considered* disvaluable, there could be valuable elements to the desire, such as pleasure at its satisfaction, but nonetheless the disvalue of the desire outweighs its value. For the purposes of this article, we bracket the issue of how to distinguish in a principled way between valuable and disvaluable mental states.

13. A sufficient condition for a desire to be disvaluable is if the satisfaction of that desire is morally impermissible. For example, the desire to torture kittens is generally disvaluable because the action of torturing kittens is generally considered to be morally impermissible.

14. One of us has provided a tentative argument to explain this difference. We propose that "autonomy of thought—understood as the condition of forming and sustaining one's desires autonomously—also possesses only conditional value. In particular, we propose that autonomy of thought lacks non-instrumental value when employed to form or sustain desires that are all things considered disvaluable" (Birks and Douglas 2017, 427–28). This means that it is not necessarily pro tanto wrong to interfere with a person's disvaluable desires. This is based on Joseph Raz's argument for the conditional value of autonomy. According to Raz, "Autonomous life is valuable only if it is spent in the pursuit of acceptable and valuable projects and relationships" (Raz 1986, 313–20).

15. Objectionable here could mean both wrong because it is harmful and wrong without harm.

interference is more morally objectionable than the former.

While the Constrained Question concerns the permissibility of interference with only C's disvaluable mental states, the Expansive Question addresses the permissibility of interfering with both valuable and disvaluable mental states.[16] Recall that the Expansive Question asks whether it is permissible to administer N on C if it has effects X, Y, and Z on C, where only effect X relates to C's offending, and effects Y, and Z do not relate to C's offending. Suppose we administer a mandatory neurointervention on an offender as a response to a violent offense. This has effect X on C, in that it makes C unlikely to commit the offense by diminishing his aggression. For the sake of simplicity, let us assume that it has two other side effects, namely, it also has effect Y on C, which means that C no longer desires to have sex with other consenting adults, and effect Z, which means that C suffers from restlessness. Even if we thought that it is not harm for C to have his mental state regarding X altered, the fact that N also causes effect Y and Z is significantly harmful. It is a qualitatively different harm from that caused by incarceration, and this is why neurointerventions are morally objectionable in a respect that incarceration is not.

It could be objected that reducing the likelihood of reoffending is distinct from reducing the likelihood of the offender having a disvaluable mental state. Therefore, we could be incorrect to claim that the difference between the Expansive Question and the Constrained Question is that the latter is concerned only with disvaluable mental states, whereas the former is concerned with both valuable and disvaluable mental states. Now, it is true that sometimes even in cases where the law is just and reasonable, criminal offenses do not involve disvaluable mental states. Consider the law against stealing, for example. Whether the desire to steal is valuable or disvaluable depends on the specific context. If one could save the life of an innocent person by stealing, then it is plausible to hold that the desire to steal could be a valuable mental state. If the desire to steal is instrumental to the desire to fund one's smoking habit, then the desire is a disvaluable mental state if the stealing is disvaluable and the instrumental benefit of smoking is slight. This is a fair objection, but for the purposes of the article we assume that the Constrained Question is narrowly focused to the extent that the neurointerventions are only concerned with reducing reoffending in cases where the desire to offend is disvaluable, and the law is just and reasonable. We accept that it might be implausible to think a neurointervention could ever have such a narrow effect. Indeed, it is further grounds for doubting that it is correct for the literature to be focused on addressing the Constrained Question. However, it is in keeping with much of the literature on the topic, which assumes for the sake of argument that wrongdoing and offending are necessarily interlinked, even when in reality they often come apart.[17]

16. We assume for the sake of argument that offenses necessarily involve only disvaluable acts.

WHAT'S SO SPECIAL ABOUT NEUROINTERVENTIONS?

The view that neurointerventions inflict an objectionable and qualitatively different harm from incarceration faces a serious objection. Suppose we impose minimally decent incarceration (I) on C as a response to a sexual offense. The diminution of liberty makes C unlikely to commit the offense, either because it prevents C from offending, or it morally educates C so that he no longer offends (Hampton 1984). However, having one's liberty diminished also has the effects of Y and Z on C. If our claim is that the harm due to effects Y and Z is objectionable in N, why are the same harmful effects, experienced to the same extent in I, different?[18]

We can state this objection clearly with the following:

> *Punishment Equivalence Thesis*: Harms to mental integrity (M) caused by a mandatory neurointervention (N) can also be inflicted by minimally decent incarceration (I). If punishment N is morally objectionable in one respect (R) because M occurs, then punishment I is equally objectionable in respect R when M occurs to the same extent as in N.

We reject the Punishment Equivalence Thesis. Even if an M occurs in N to the same extent as in I, it does not follow that N and I are equally objectionable in that same respect. The harms to mental integrity caused by N could be morally more problematic than identical harms caused by I. The reason, we explain, involves the fact that at least some Ms caused by N are intended, whereas the same Ms caused by I could be unintended but foreseen.

This matters if one holds that it can be more difficult to justify intended harm than harm that is unintended but foreseen.[19] If one accepts this distinction, and some

17. For an illuminating discussion on the relationship between wronging and offending see Tadros (2016).

18. It has been suggested that it is unrealistic to think that incarceration would cause the same harmful mental effects to the same extent as a neurointervention. We agree that it might be true that in practice the mental effects of incarceration could be less harmful. However, if our argument is able to contend with the stronger objection, that the same mental effects are experienced to the same extent in incarceration, then our argument would also be successful for cases where the mental effects is experienced to a lesser extent.

19. We do not attempt to defend this commonly held view here. For a paper setting out the implications for denying intentions are relevant for permissibility, see Husak (2009).

Ms caused by N are intended, whereas the same Ms caused by I are unintended but foreseen, this would be grounds to reject the Punishment Equivalence Thesis.[20] The two punishments are not necessarily equally objectionable in one respect when the same Ms occur to the same extent in both cases.[21] This is why the Expansive Question reveals a moral problem that is absent when we address the Constrained Question. Once we consider the other effects of neurointerventions beyond making the offender less likely to offend, there can be a difference in intention between harms caused by the two types of punishment.

This might appear initially puzzling. How could some Ms be intended in N but the same Ms not intended in I? A proponent of administering mandatory neurointerventions might say, it is unfortunate that they result in harms M due to effects Y and Z, but he only wants to cause effect X, and the other effects Y and Z are unintended. He would still want to administer the neurointervention, even if these other effects did not occur. He might even wish that scientific advances in neurointerventions mean that only effect X occurs when administering the neurointervention. It could be asked, isn't this sufficient to show that these other harms are unintended, and so equally difficult to justify as the same harms in I? In short, no, it is not.

There is a voluminous literature discussing how one can draw a distinction between harms that are intended, and those that are unintended but foreseen. Indeed, it is famously objected that the distinction between what a person intends and what he foresees is so arbitrary, it renders the distinction meaningless (Foot 2002, 21). This is known as the problem of closeness, and it can be illustrated by the following classic case:[22]

> *Cave:* A number of explorers are trying to escape an underground cave from rapidly rising water. The largest member of their group tries to escape first, but gets stuck in a small hole, the only exit to the cave. The explorers unsuccessfully try to dislodge the stuck explorer. The only way to escape and avoid drowning is to use dynamite to blow up the stuck explorer.

Let's set aside whether or not it is permissible to blow up the stuck explorer, and instead focus on their intention. It is accepted that if the explorers were to be able to claim that they did not intend to kill the large man, but only intended to blow him up as means of escaping, then this would render the distinction between intended and merely foreseen harms meaningless. The blowing up of the stuck explorer is too close to his death for it to be plausible to claim that his death is an unintended but foreseen consequence of blowing him up. In order for it to have any moral import, it needs to provide a less arbitrary distinction than this (Foot 2002, 21).

There have been many attempts to draw a satisfactory distinction between intended and merely foreseen harms.[23] We cannot provide an overview of the various distinctions here.[24] Rather, we focus on what we think is the most successful attempt, namely, William Fitzpatrick's account of relations between states of affairs (Fitzpatrick 2006). If a relation between two states of

20. It might even be thought that a person's interest in mental integrity can be harmed only if his desires are created or altered intentionally. Indeed, it would be implausible to hold that unintentionally creating or altering a person's desire is always a harm to a person's interest in mental integrity. Our behavior frequently creates and alters the desires of others, and without some form of an intentionality requirement, a doubtfully large array of actions would be rendered wrong. For example, imagine we paint our house green because it is our favourite color. As it happens, the color green has a soothing effect and reduces our neighbour's desire to be aggressive. It seems implausible that this harms our neighbor's mental integrity, despite the fact that it alters his desires. However, we might think that there is a moral difference if we were to paint our house green with the intention of changing our neighbor's desires. We do not need to hold this strong view here in order for our argument to be successful. For an argument doubting that there is a morally relevant difference between intentionally altering a desire by administering a neurointervention and by painting a prison wall green, see Douglas (2018).

21. Some proponents of the importance of intentions to permissibility hold that only some intended harms are more difficult to justify than if they were unintended but foreseen. For example, it might be thought that there is a rights restriction, namely, the view that only harms relating to a rights violation regardless of the intention of the agent are more difficult to justify when intended rather than foreseen. We do not need to take a stand on this issue here, but instead we assume that the harm of interfering with a person's mental integrity is the type of harm that is more difficult to justify when intended rather than merely foreseen. For a clear discussion of this, see McMahan (1994) and Husak (2009).

22. This case is adapted from Foot (2002, 21).

23. For a critical overview of the many attempts, see Nelkin and Rickless (2015). Some philosophers have argued that we can make sense of the role of intentions to permissibility without needing to solve the problem of closeness. For example, Victor Tadros argues that "What matters is whether a person executes an intention to affect others in a way that foreseeably harms them" (Tadros 2015, 74). See also: Warren Quinn (1993). It is beyond the scope of this article to discuss the implication of this view for harms of mandatory neurointerventions.

24. It is possible that our argument would be successful if one employed an alternative solution to the problem of closeness, such as Bennett's conceivability requirement (Bennett 1995, 213). However, we are unable to provide an overview of the many attempts to solve the problem of closeness and consider whether each one can show that the harms of mental integrity inflicted by neurointerventions are intended whereas the same effects are merely foreseen in incarceration. We believe that Fitzpatrick's account is the most compelling solution to the problem of closeness, and it also has these implications for distinguishing the harms to mental integrity of neurointerventions from incarceration.

affairs is known by an agent to be constitutive rather than causal, then the agent cannot claim that one is unintended but merely foreseen (Fitzpatrick 2006, 206). However, if the relation between states of affairs is causal, then it could be unintended, and merely foreseen.[25]

For example, in Cave, the state of affairs of blowing up the stuck explorer is known by the agents to be constitutive of the state of affairs of killing him, so it is not possible for them to intend to do one without intending the other. The fact that the explorers might say that they would rather their fellow explorer did not die, or that they wish there were a way to piece him back together, does not mean that killing him is unintended. The fact that they knew the state of affairs of blowing him up is constitutive of the state of affairs of his death means that his death is intended, regardless of whether they regret it. As Fitzpatrick says, "One can no more aim at a man's being blown to bits without aiming at his being killed than one can aim (literally) at a spot on a target without aiming at the target it partly constitutes" (Fitzpatrick 2006, 595).

It might be helpful to further clarify this distinction by looking at the classic Trolley case:

> *Trolley:* A trolley is hurtling down a track toward five people. It will kill all five. You are standing alongside a lever that will divert the trolley down a different track that will kill one person.[26]

If we turned the lever to direct the trolley to kill one person rather than the five, we do not intend to kill the one person, but rather, his death is only causally related to turning the trolley. Turning the trolley causes the person to die, but the relation between the state of affairs of turning the trolley and killing the one man is not constitutive; rather, it is causal. As a result, we are able to say intelligibly that we did not intend to kill the person on the track, even though when we pull the lever his death is foreseen. This can be further emphasized by looking at the following familiar case:

> *Large Man:* A trolley is hurtling down a track toward five people. It will kill all five. You are standing alongside a large man such that if you push him, he will fall and block the trolley, stopping it from killing the five on the track, but die in the process.[27]

In this case, the state of affairs of pushing the large man to block the trolley is known by the agent to be constitutive of the state of affairs of the large man falling to his death, so the agent cannot say that killing the large man is an unintended but foreseen consequence of pushing him from the bridge.[28]

Before we proceed to relate the constitutive/causal distinction to punishment, we should make a couple of further clarifications to this relation between states of affairs. One way Fitzpatrick does this is to explain what the constitutive relation is not. First, it is not the fact that one state of affairs will necessarily occur following the other. Fitzpatrick is clear that he rejects what is sometimes called a causal necessitation principle (Bennett 1995, 209). This captures too much. After all, in Trolley, it might be the case that given the speed and size of the trolley, pulling the lever to redirect the trolley will necessarily cause the death of the one person. Yet we should still be able to say that killing the one is merely foreseen but unintended. The constitutive/causal relation is also distinct from logical entailment, which captures too little. In Cave, blowing up the large man into pieces does not logically entail his death. After all, it is logically possible to blow up a person without killing him. For instance, technology might exist in the future to put the person back together without resulting in his death. In summary, then, the constitutive relation between states of affairs is stronger than causal necessitation but weaker than logical entailment.

With these further clarifications in hand, we can now explain how the relation between the state of affairs of the administering the neurointervention and the state of affairs of Ms caused by effects Y and Z are constitutive rather than causal. We consider two different forms of neurointervention. We first consider a simple example, where the neurointervention causes two effects of the same type, and then proceed to discuss a more complex case, where the neurointervention causes two different effects.

Suppose in order to make C less likely to commit a sexual offense, we administer an antilibidinal pharmacological agent N that decreases C's testosterone. When committing the offense, and before he is punished, C has testosterone to the score of 50. Following administering N, C has testosterone to the score of 20, which means he no longer possess a disvaluable sexual desire. However, when a person's testosterone is diminished to 20 it also has the effect that he is less likely to be able to have any form of valuable sexual desire. Given that our action of administering N diminishes testosterone to 20, and this decrease interferes with both valuable and disvaluable mental states, we cannot claim only one effect is

25. For the sake of simplicity, we omit one important aspect of Fitzpatrick's account, namely, that the relation between state of affairs needs to be natural, rather than conventional. That does not affect our argument here.
26. This famous case is adapted from Judith J. Thomson (1985, 1397). For the original Trolley Problem see Philippa Foot (2002).
27. This case is also adapted from Thomson (1985, 1409).
28. It could be questioned whether Fitzpatrick is correct that these are constitutive relations, rather than causal. We proceed with the assumption that the relationship between the state of affairs in Cave and Large Man is constitutive. We do not defend this assumption here. Our point is that if these relations are constitutive, then so are the relations between administering the neurointervention and the harms M due to effects Y and Z.

unintended but foreseen. The state of affairs of diminishing C's testosterone by administering N is constitutive of the state of affairs of C being less likely to have a valuable sexual desire. We cannot say that we didn't intend to make C less likely to have a valuable sexual desire, and only eliminate C's disvaluable sexual desire.

In contrast, suppose that in order to make C less likely to commit a sexual offense we diminish C's liberty by deploying minimally decent incarceration. The diminution of liberty also has the effect of making C less likely to have a valuable sexual desire. The relationship is causal, as although the diminution of liberty causes C not to have a valuable sexual desire, the diminution of C's liberty isn't constitutive of eliminating his valuable sexual desire. We can distinguish between the diminution of the offender's liberty, and the respective results of doing this, such as a reduction in his valuable sexual desire. If there is conceptual space to say that an agent can aim at diverting a trolley but not aim at a person being killed, similarly we can state that we can aim to diminish the offender's liberty but not aim to eliminate the offender's valuable sexual desire.

Admittedly, there are a number of different types of neurointerventions, and the use of antilibidinal pharmacological agents might be thought to be an easier case for us. It might be easier because the neurointervention has only one type of effect, namely, it diminishes sexual desires, both valuable and disvaluable. It might be thought that our task becomes more complicated when we consider a neurointervention that has more than one type of effect on the offender.[29] For example, one proposed possible future neurointervention will diminish aggression by increasing the offender's serotonin, but this will also have quite different effects, such as diminishing valuable sexual desires, or causing restlessness.[30]

However, even when an intervention has two quite different effects, it does not follow that they cannot be both constitutive of the intervention. For example, imagine that in order to get A to stop talking, B shoots his shotgun at A's brainstem. The shotgun shell has the effect of stopping A talking. But in addition to the effect of stopping A talking, shooting A's brainstem also stops A breathing too. B cannot claim that he only intended to stop A talking by shooting A's brainstem, when he knows the relation between the states of affairs of shooting the brainstem and stopping A breathing are constitutively related. B cannot claim that he did not intend to stop A from breathing, and only intended to stop A from talking.

Similarly, let's consider a neurointerventions that has two or more different types of effects. Suppose we increase C's serotonin with a neurointervention in order to make him less aggressive. Serotonin has multiple functions. Further suppose, for the sake of simplicity, that C when committing the offense, and before he is punished, has serotonin to the score of 20. When C's serotonin is at 20, C is likely to be aggressive. Following the serotonin neurointervention, C has serotonin to the score of 50, which means he is less likely to be aggressive. However, when C's serotonin is increased to 50 this makes him suffer from restlessness. Given that our action of administering N increases C's serotonin to 50, and this increase has two effects, we cannot claim one of the effects is unintended but foreseen. We cannot say that we did not intend to make C suffer restlessness. The state of affairs of increasing serotonin by administering N is constitutive of the state of affairs of C suffering from restlessness.[31]

Compare this to the relation of harms to mental integrity caused by deploying minimally decent incarceration on C. The harm of having one's liberty diminished is intended. However, the state of affairs of diminishing C's liberty is only causally related to the state of affairs where C experiences the harm of restlessness. Note that even though the harms are less likely to occur in I than in N, this is a different relation between states of affairs, and this is the case even if the same harms do occur to the same extent. Hence we can say that the harm of restlessness caused by the effects in minimally decent incarceration could be unintended but foreseen, but the identical harms, experienced to the same extent, are intended in N. Thus, we should reject the Punishment Equivalence Thesis.

29. It is made particularly difficult to set out the distinction between constitutive and causally related states of affairs given that Fitzpatrick does not provide a set of necessary and sufficient conditions for states of affairs to be constitutively related. In a critical discussion of Fitzpatrick's solution to the problem of closeness, Nelkin and Rickless (2015) consider the reasons for this omission. They write, "It is unclear whether Fitzpatrick thinks that the constitution relation is not susceptible of definition (because, say, it is a prototype concept or family resemblance concept) or whether he thinks that the constitution relation is definable but the details of the definition unnecessary or distracting" (390).

30. These are the side effects of selective serotonin reuptake inhibitors (SSRIs) deployed to increase serotonin. For an analysis of the relationship between serotonin and aggression see Duke et al. (2013). It is important to note that our understanding of the relationship between serotonin and agression is far from complete. See Crockett (2014).

31. We note that the option to treat the side effect of restlessness with, say, a benzodiazepam such as Xanax does not change that the fact that the relation between states of affairs is constitutive. Recall the case of Cave, where blowing up the large man does not logically entail his death, because technology might exist in the future to put the person back together without resulting in his death. The existence of the technology does not mean the relation between the state of affairs of blowing up the stuck explorer and the state of affairs of killing him is not constitutive. The fact that we could eliminate the restlessness with Xanax would be the equivalent of putting back together the blown-up explorer—but would not change the relation between the states of affairs from being constitutive.

It is worth repeating that this account is only claiming that if the relation between states of affairs is causal, then it *could* be unintended, and merely foreseen. But there are cases where the relation between states of affairs is causal and yet still are intended by the agent. For example, in the case of incarceration, the relationship between the state of affairs of diminishing the offender's liberty and its mental effects is causal. However, bringing about the mental effects by diminishing the offender's liberty can still be intended. Suppose a vindictive judge knew that administering a neurointervention would be impermissible, but she wants to inflict certain mental effects on the offender. In order to do this, she imposes incarceration on the offender in order to achieve the same mental effects as the neurointervention. In such a case, despite the relation being causal rather than constitutive, the mental effects are intentionally inflicted on the offender.

It could be argued that in the distant future, neurointerventions will exist that will not affect an offender's valuable mental states, because our knowledge of the brain's function will have dramatically increased. Perhaps interventions could be developed that are so specific they only produce one effect related to the offender's disvaluable mental states.[32] However, this is not a relevant consideration to the relation between states of affairs now. Consider the following: In the future, it might be possible to have some substance that would temporarily shrink the stuck explorer in Cave so that his fellow explorers could escape, but at present, that substance can only shrink a person and kill him. We wouldn't say that the fact that the shrinking substance might be developed in the future so that it didn't kill its intervenee means that shrinking the explorer now is not constitutive of killing the explorer. The same holds for neurointerventions. Just because there is a possibility that in the future interventions might be developed that have no effects besides interfering with disvaluable mental states, this does not mean that the interferences with valuable mental states due to current neurointerventions are not constitutive of administering the neurointerventions.[33] The possibility of this in the future does not change the relation between the states of affairs now.[34]

CONCLUSION

Neurointerventions interfere with an offender's mental states. The interference will not only affect his disvaluable mental states, but will also have wide-reaching implications for his valuable mental states. In this article, we have suggested that this has both theoretical and practical significance for the permissibility of administering mandatory neurointerventions. We considered the objection that incarceration also involves an equally objectionable interference with the offender's valuable mental states. We argued that even if it were the case that the offender is harmed to the same extent in the same respect, it does not follow that the harms are morally equivalent. If one holds that intended harm is more difficult to justify than harm that is unintended but merely foreseen, this means neurointerventions could be morally objectionable in a significant respect that incarceration is not.

ACKNOWLEDGMENTS

We thank the following, who provided helpful comments and objections to the article: Christopher Bennett, Ian Carroll, Tom Douglas, Sam Kiss, Michael Lundie, Duncan Purves, Lorenzo Del Savio, Tony Taylor, Frej Klem Thomsen, and Pete West-Oram. We also thank audiences at the following events: Braga Meetings in Ethics and Political Philosophy at the University of Minho, the 10th International Conference on Applied Ethics at Hokkaido University, Sapporo, and the Rocky Mountains Ethics Congress, University of Colorado at Boulder. We are particularly grateful to Robert Kelly for his insightful response to an earlier version of the article at the latter event. We also thank two anonymous reviewers for the perceptive objections and suggestions. Finally, we thank Mona Rudolf for her invaluable research assistance.

FUNDING

This article was written while supported by the German Research Foundation (BU 2450/2-1).■

32. It is important to emphasize, as mentioned earlier, that this is highly unlikely and would require us to reject a fundamental principle of pharmacology.

33. Similarly, in the distant future, we might imagine that technology may have developed so that there exists an incredibly precise ray gun that is able to shoot a laser beam that temporarily stops a person talking, without any other effects. Nevertheless, the fact such technology might exist in the future does not change whether B intends to stop A breathing or not.

34. It is worth noting that the constitutive/causal distinction has a significant implication for a common justification of one type of euthanasia. It is often proposed that it is permissible to administer a drug with the intention to relieve a patient's pain, even if the doctor foresees that the drug will hasten the patient's death. In contrast, if the doctor injected the same drug with the intention to kill the patient, this would be impermissible. If one accepts the constitutive/casual distinction, however, such a move is not available. A doctor who administers a drug in order to relieve pain while foreseeing that it will hasten the death of the patient cannot say that he did not intend to hasten the patient's death. The state of affairs of injecting the drug is known by the doctor to be constitutive of the state of affairs of bringing about a patient's death, so it is not possible for him to intend to do one without intending the other. In fact, one implication of this account is that whenever a drug is administered to a patient with the knowledge that it is constitutive of the state of affairs of the patient experiencing side effects, the side effects cannot be unintended. It is only if the doctor is unaware that the drug has a particular side effect that could it not be intended.

ORCID

David Birks http://orcid.org/0000-0002-8416-1818
Alena Buyx http://orcid.org/0000-0002-5726-7633

REFERENCES

Adebowale, V. 2010. Diversion, not detention. *Public Policy Research* 17(2): 71–74.

Bennett, C. 2018. Intrusive intervention and opacity respect. In *Treatment for crime: Philosophical essays on neurointerventions in criminal justice*, ed. D. Birks and T. Douglas, Oxford: Oxford University Press.

Bennett, J. 1995. *The act itself*. Oxford: Clarendon Press.

Berman, G., and A. Dar. 2016. *Prison population statistics*. London: House of Commons Library.

Birks, D., and T. Douglas. 2017. Two ways to frustrate a desire. *The Journal of Value Inquiry* 51(3): 417–434.

Boonin, D. 2008. *The problem of punishment*. Cambridge: Cambridge University Press.

Briken, P., and M. P. Kafka. 2007. Pharmacological treatments for paraphilic patients and sexual offenders. *Curr Opin Psychiatry* 20(6): 609–613. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=med5&NEWS=N&AN=179217 64. Accessed 19 July 2016.

Bublitz, C. J., and R. Merkel. 2014. Crimes against minds: On mental manipulations, harms and a human right to mental self-Determination. *Criminal Law and Philosophy* 8(1): 51–77.

Bureau of Justice Statistics. 2009. Key Statistics: Estimated number of persons under correctional supervision in the United States, 1980-2014. www.bjs.gov. Accessed 19 July 2016.

Christman, J. 1991. Autonomy and personal history. *Canadian Journal of Philosophy* 21(1): 1–24.

Conner, M. W., D.-C. Catherine., and L. C. Green, et al. 2012. Drug toxicity. In *Principles of pharmacology: The pathophysiologic basis of drug therapy*, ed. D. E. Golan and A. H. Tashjian Philadelphia: Wolters Kluwer Health.

Crockett, M. J. 2014. Moral bioenhancement: A neuroscientific perspective. *Journal of Medical Ethics* 40(6): 370–371.

Douglas, T. 2014. Criminal rehabilitation through medical intervention: Moral liability and the right to bodily integrity. *The Journal of Ethics* 18(2): 101–122.

Douglas, T. 2018. Neural and environmental modulation of motivation: What's the moral difference? In *Treatment for crime: Philosophical essays on neurointerventions in criminal justice*, ed. D. Birks and T. Douglas, Oxford: Oxford University Press.

Douglas, T., and D. Birks. 2018. Introduction. In *Treatment for crime: Philosophical essays on neurointerventions in criminal justice*, ed. D. Birks and T. Douglas, Oxford: Oxford University Press.

Duggan, C. 2011. Dangerous and severe personality disorder. *The British Journal of Psychiatry: The Journal of Mental Science* 198(6): 431–433.

Duke, A., L. Bègue, R. Bell, and T. Eisenlohr-Moul. 2013. Revisiting the serotonin-Aggression Relation in humans: A Meta-Analysis. *Psychological Bulletin* 139(5): 1148–1172. http://www.ncbi.nlm.nih.gov/pubmed/23379963. Accessed 19 July 2016.

Durose, M. R., A. D. Cooper, and H. N. Snyder. 2014. Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010. Bureau of Justice Statistics: Special Report. NCJ 244205.

Ekstrom, L. W. 2005. *Autonomy and personal integration in James Stacey Taylor, ed., Personal Autonomy*. Cambridge: Cambridge University Press.

Fitzpatrick, W. J. 2006. The intend/foresee distinction and the problem of 'closeness. *Philosophical Studies* 128(3): 585–617.

Foot, P. 2002. *Virtue and vices and other essays in moral philosophy*. Oxford: Clarendon Press.

Golan, D. E., and A. H. Tashjian. 2012. *Principles of pharmacology: The pathophysiologic basis of drug therapy*. Philadelphia: Wolters Kluwer Health.

Hampton, J. 1984. The moral education theory of punishment. *Philosophy and Public Affairs* 13(3): 208–238.

Haney, C. 2003. The psychological impact of incarceration: Implications for post-prison adjustment. In *Prisoners once removed: The impact of incarceration and reentry on children, families, and communities*, ed. J. Travis and M. Waul, 33–66. Washington, D.C.: Urban Institute Press.

Henrichson, C., and R. Delaney. 2012. *The price of prisons: What incarceration costs taxpayers*. New York, NT: Vera Institute of Justice.

Husak, D. 2009. The costs to criminal theory of supposing that intentions are irrelevant to permissibility. *Criminal Law and Philosophy* 3(1): 51–70.

James, N. 2013. The federal prison population buildup: Options for congress. *Congressional Research Service* 42937: 7–5700. http://www.prisonstudies.org/highest-to-lowest/occupancy-level?field_region_taxonomy_tid=All. Accessed 30 May 2016.

Kane, S. 2013. Interview with the most rehabilitated prisoner in america, ann arbor. *Litigation* 39(3): 16–23.

Krakowski, M. I., and P. Czobor. 2014. Depression and impulsivity as pathways to violence: Implications for antiaggressive treatment. *Schizophrenia Bulletin* 40(4): 886.

Levy, N. 2007. *Neuroethics: Challenges for the 21st century*. Cambridge: Cambridge University Press.

Lippert-Rasmussen, K. 2018. *The self-Ownership Trilemma, extended minds, and neurointerventions in treatment for crime: Philosophical essays on neurointerventions in criminal justice*, ed. D. Birks and T. Douglas, Oxford: Oxford University Press.

McMahan, J. 1994. Revising the doctrine of double effect. *Journal of Applied Philosophy* 11(2): 201–212.

McMahan, J. 2018. Moral Liability to 'Crime-Preventing Neurointervention". In *Treatment for crime: Philosophical essays on neurointerventions in criminal justice.*, ed. D. Birks and T. Douglas, Oxford: Oxford University Press.

Nelkin, D. K., and S. C. Rickless. 2015. So close, yet so far: Why solutions to the closeness problem for the doctrine of double effect fall short. *NOUS* 49(2): 376–409.

Oshana, M. 2006. *Personal autonomy in society*. Aldershot: Ashgate Press.

Petersen, T. S., and K. Kragh. 2017. Should violent offenders be forced to undergo neurotechnological treatment? A critical discussion of the 'freedom of thought' objection. *Journal of Medical Ethics* 43(1): 30–34.

Quinn, W. 1993. *Actions, intentions and consequences: The doctrine of double effect in morality and action*. Cambridge: Cambridge University Press.

Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.

Ryberg, J. 2012. Punishment, pharmacological treatment, and early release. *International Journal of Applied Philosophy* 26(2): 231–44.

Shaw, E. 2014. Direct brain interventions and responsibility enhancement. *Criminal Law and Philosophy* 8(1): 1–20.

Shaw, E. 2018. Against the mandatory use of neurointerventions in criminal sentencing. In *Treatment for crime: Philosophical essays on neurointerventions in criminal justice*, ed. D. Birks and T. Douglas Oxford: Oxford University Press.

Sitaram, R., A. Caria, and R. Veit, et al. 2014. Volitional control of the anterior insula in criminal psychopaths using real-time fMRI neurofeedback: A pilot study. *Frontiers in Behavioral Neuroscience* 8: 344. doi:10.3389/fnbeh.2014.00344.

Smith, R., R. Grimshaw, R. Romeo, and M. Knapp. 2017. Poverty and Disadvantage among Prisoners' Families. Joseph Rowntree Foundation.

Tadros, V. 2015. Wrongful intentions without closeness. *Philosophy & Public Affairs* 43(1): 52–74.

Tadros, V. 2016. *Wrongs and crimes*. Oxford: Oxford University Press.

Thomson, J. J. 1985. The trolley problem. *The Yale Law Journal* 94(6): 1395–415.

Vallentyne, P. 2018. Neurointerventions, self-ownership, and enforcement rights. In *Treatment for crime: Philosophical essays on neurointerventions in criminal justice*, ed. D. Birks and T. Douglas, Oxford: Oxford University Press.