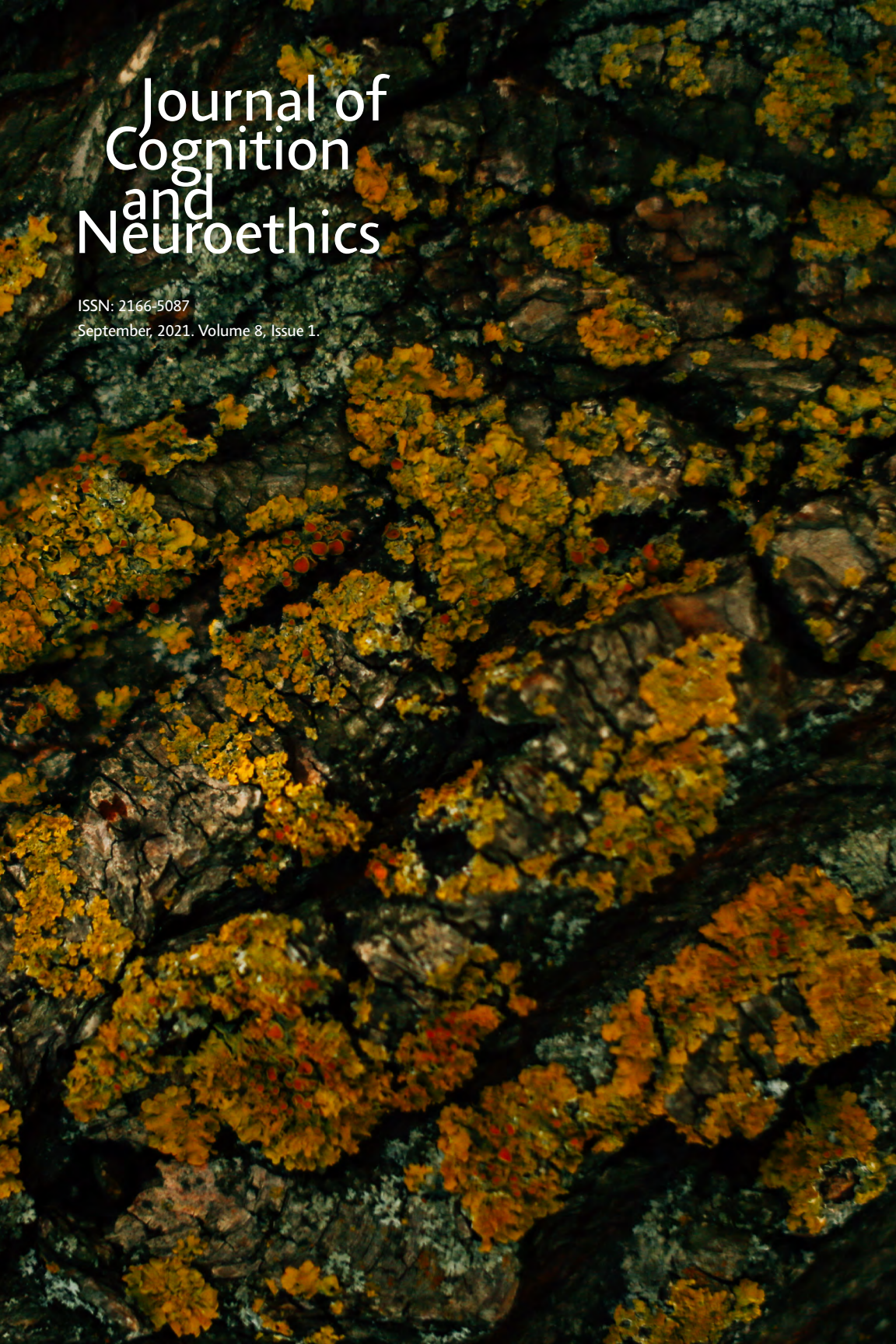
Journal of Cognition and Neuroethics

ISSN: 2166-5087
September, 2021. Volume 8, Issue 1.



Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Production Editor
Zea Miller

Publication Details
Volume 8, Issue 1 was digitally published in September of 2021 from Flint, Michigan, under ISSN 2166-5087.

The *Journal of Cognition and Neuroethics* is produced by the Center for Cognition and Neuroethics. For more on CCN or this journal, please visit cognethic.org.

Center for Cognition and Neuroethics
University of Michigan-Flint
Philosophy Department
544 French Hall
303 East Kearsley Street
Flint, MI 48502-1950

Table of Contents

Journal of Cognition and Neuroethics

Interventionist Advisory Brain Devices, Aggression, and Crime Prevention

Sebastian Jon Holmen
Roskilde University

Jesper Ryberg
Roskilde University

Biographies
Sebastian Jon Holmen (ORCID 0000-0003-2774-941X) is a PhD research fellow at the Department of Philosophy and Science Studies at Roskilde University.

Jesper Ryberg is Professor of Ethics and Law at Roskilde University.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). September, 2021. Volume 8, Issue 1.

Citation
Holmen, Sebastian Jon, and Jesper Ryberg. 2021. "Interventionist Advisory Brain Devices, Aggression, and Crime Prevention." *Journal of Cognition and Neuroethics* 8 (1): 1–22.

Interventionist Advisory Brain Devices, Aggression, and Crime Prevention

Sebastian Jon Holmen and Jesper Ryberg

Abstract

Some novel brain devices are able to predict neural events, making it possible for the device to advise its user to engage in the appropriate countermeasures before the event takes place. Other devices can automatically discharge such countermeasures on its user's behalf. In this paper, we consider some of the ethically questions that will arise if it becomes possible to combine such advisory and interventionist capabilities in a brain device to combat episodes of uncontrollable impulsive aggression. Specifically, if a device becomes available that can monitor and collect an offender's neural data, give him behavioural advice based on this data, and discharge countermeasures *unless* the offender actively keeps it from doing so, should such an interventionist advisory brain device be mandated to some offenders? In the following, we critically examine a range of plausible reasons to oppose such use related respectively to the device's capacity to monitor and collect an offender's brain data, its advice-giving feature, and its ability to discharge aggression-hampering treatment absent offenders' active dissent. We find that, surprisingly, none of the considered reasons can stand further scrutiny.

Keywords

Interventionist Advisory Brain Devices, Crime prevention, Explosive Aggression, Neurointerventions

1. Introduction

The ability of some novel brain devices to predict the occurrence of specific neural events, providing patients with the possibility of engaging in the necessary countermeasures or for an automated therapeutic activation system[1] to do so on their behalf, will potentially transform the way neurological disorders and neurodegenerative diseases will be treated in the future. For example, some devices tested on patients suffering from epileptic seizures use trained algorithms to detect the neuronal patterns related to a seizure and inform the patient through a handheld device how likely it is to materialize. This makes it possible for these patients to avoid many of these seizures

1. Such an automated system could, but need not be, similar to novel methods of brain stimulation in which an implant detects brain patterns related to the condition the implant is intended to counteract and automatically adjust the timing, location, and intensity of the delivered stimulation in response to these data (Malekmohammadi et al. 2016; Glannon and Ineichen 2016).

by pre-emptively administering the appropriate medication (Cook et al. 2013). It has been speculated that the ability of such advisory brain devices to forecast neural events may perhaps be employed in the future to prevent socially undesirable behaviour, such as uncontrollable aggressive and violent behaviour, which is similarly preceded by predictable neural changes (Gilbert 2015). In this paper, we will consider some of the ethical questions the development of such a device might raise.

Suppose that an advisory brain device is developed that is able to predict upcoming episodes of explosive aggression in individuals unable to control such violent impulses if they materialize, and could advise them to take the appropriate countermeasures (e.g., instruct them to leave a high-risk situation or simply to calm down) within a specified timeframe. Suppose, further, that this device had the additional feature of automatically being able to discharge electrical stimulation or medication to keep these explosive episodes from materializing *unless* the offender actively keeps it from doing so (e.g., by remote control) within the specified timeframe.[2] Finally, suppose that employing such *interventionist advisory brain devices* were effective in reducing re-offending rates among offenders suffering from such otherwise uncontrollable episodes of aggression, without them having to endure serious side-effects from being connected to the device or from its continued functioning. Should such devices then be mandated by the state to this group of offenders? Although other questions related to the ethics of employing brain devices and other neurointerventions on offenders are increasingly being addressed in the literature (e.g., Ryberg 2020; Birks and Douglas 2018; Douglas 2014; Ryberg 2012; Bublitz and Merkel 2014; Shaw 2014; Holmen 2020; Kirchmair 2019; Petersen and Kragh 2017; Ligthart et al. Forthcoming), the possibility of combining the advisory and interventionist capabilities of brain devices to combat recidivism has received virtually no scholarly attention.[3]

On the one hand, it seems clear that if such interventionist advisory brain devices prove effective in reducing re-offending, the possibility of preventing grievous harm to victims of future crimes of aggression (and their families) provides a strong moral

2. Either the offender could be informed in advance of the timeframe between the point at which the first piece of advice is given and the point at which the drug is being automatically discharged, or the action-guiding advice could be supplied via information to the carrier that the treatment will be initiated within a timeframe (for instance, the offender could receive a message along the lines: "Treatment will be initiate within one minute if the deactivation button is not pushed"). In the following, we will not engage in detailed speculation of what will constitute the optimal design of the device.

3. The only ethical attention this possibility seems to have attracted is contained in a short comment from Jesper Ryberg (2015).

reason to employ them. Further, the offender himself (and his family) will be spared the deprivations that re-punishing the offender entails. Finally, and more generally, the resources that would otherwise be spent on re-punishing these offenders could be allocated to other (perhaps more morally desirable) projects. On the other hand, several important moral objections may be raised in response to a proposal to mandate these devices to some offenders. Such objections may relate to one or more of the three features we are here proposing the interventionist advisory device would have: (1) its monitoring and collection of information about an offender's neural environment; (2) its advice-giving feature; and (3) its ability to discharge countermeasures unless the offender actively keeps it from doing so. In the following sections, we will critically examine several plausible objections that could be raised in relation to each of these features. Specifically, the paper proceeds as follows. In Section 2, we critically examine whether the fact that employing the device involves the non-consensual monitoring and collection of information about parts of an offender's neural environment constitutes a plausible reason to oppose its use. Section 3 discusses whether it would be convincing to oppose the use of the device because it may present advice to the offender against his will. We shall then, in Section 4, discuss the intuitively plausible view that it is morally significant that the stimulation or medication delivered by the device would not be administered as the result of an active decision on the part of the offender, but would rather come about due to an act of omission. In Section 5, we consider whether there are good reasons for thinking it morally preferable not to prevent episodes of explosive aggression with brain stimulation or medication. Section 6 discusses whether it may be wrong to use such devices on offenders because it involves requiring them to have the device placed on their body against their will. The question of whether the potential impact of the advisory device on offenders' self-conception should lead us to judge against employing them will be confronted in Section 7. Finally, in Section 8, we summarize and conclude that, surprisingly, none of these *prima facie* plausible considerations can stand further scrutiny.

Before embarking on this discussion, however, a few comments are required regarding the scope of the paper. First, although we shall briefly comment on this question when concluding the paper, we do not discuss concerns in depth that may be raised narrowly in relation to the implementation of such a device. These will of course depend upon what the implementation requires precisely. The following discussion is nevertheless motivated by the assumption that ethical concerns regarding the mandated use of an interventionist advisory brain device will not be limited exclusively to the way such a system is implemented. Second, we do not wish to consider the potential misuse

of this device (e.g., the risk that it may be hacked).[4] As with all other types of technology, an interventionist advisory brain device may of course be vulnerable to various kinds of misuse. In the following, however, we are interested in the arguments that might be raised against the mandated use of such a device under the assumption that it works properly.

2. Mental Privacy

To be able to present an offender with behavioural advice and (if he does not instruct it not to do so) to discharge countermeasures, the interventionist advisory device would need to gather information about the offender's real-time neural environment. It might, however, be suggested that the non-consensual monitoring and collection of information about parts of an offender's inner life constitutes a case of involuntary mind-reading, and as such is a violation of his moral right to mental privacy. Several commentators have indicated, that they believe this kind of rights violation by neurotechnological means to be a matter of great ethical concern (e.g. Lavazza 2018; Ienca and Andorno 2017), a view that we shall not dispute in the present paper.[5] But should considerations regarding mental privacy lead us to reject the use of interventionist advisory devices on some offenders? In our view, one reason will suffice to show that the answer should be in the negative.

The reason is that it is plausibly not sufficient for a violation of mental privacy simply that the device monitors and collects information about the offender's neural environment. For a right to mental privacy to be violated by the device it seems, in our view, necessary that someone other than the offender should have, or gain, non-consensual access to the brain data the device collects. This is not something the device under consideration would be designed to do.[6] Rather, as described in the introduction, the device would collect and process the information for the sole purpose of being able to present the offender with behavioural advice. By, admittedly rough, analogy, a surveillance camera installed in a person's home (even one placed there without his

4. For a discussion of this possibility, see for example Pycroft et al. (2016).

5. For some issues regarding the specification of the scope of such a moral right to mental privacy, see Ryberg 2017.

6. A possibility that might make a privacy-based objection relevant would be the storing of an offender's brain data by the state, but this is not what we are considering here.

consent) can hardly be said to violate his privacy if the person being filmed is the only one presented with the information observed by the camera.[7] Thus, it seems that a concern for offenders' mental privacy – important as this may be in other contexts – is unable to block the use of the interventionist advisory devices on some offenders.

3. Unwanted Advice

If the non-consensual collection of brain data does not in itself constitute a convincing moral reason to oppose employing the advisory interventionist device on offenders, perhaps the advice-giving feature made possible by the collection of data does so. Specifically, should the fact that the advice presented to the offender – in the form of recommendations to take certain precautions – is unsolicited and might be presented to him against his will have us oppose the use of such devices? Generally speaking, there is no doubt that receiving unsolicited advice can sometimes be unwelcome to the advisee. We see no reason to think this might not also be the case with the behavioural advice provided by the interventionist advisory device. Upon further reflection, however, it is clear that the potential unwanted nature of the advice is not a convincing reason to oppose its proposed use.

First, the advice could be delivered to the offender in a way that makes it possible for him to avoid receiving it. If – as was the case with the use of advisory devices on patients with epilepsy cited in the introduction (Cook et al. 2013) – the advice is provided through a handheld device, for example, then the offender could simply place the device out of sight.

Second, even if we assume that the offender cannot avoid the advice from the device and that it is being presented to him against his will, it is hardly clear that this should raise a moral red flag. The reason is that it is very difficult to accept that it should generally be morally wrong to present someone with advice they do not want, particularly if it would be beneficial for the advisee to receive the advice.[8] For example, it is surely difficult to

7. Although the non-consensual installation of the camera in the house might of course be wrong in itself because it, e.g., violates his property rights.

8. It might perhaps be objected that, in the present case, the advice would not be of benefit to the advisee because even if he decided not to act in accordance with it – for instance, by calming down or leaving the location – the risk that he would be involved in criminal conduct would nevertheless still be prevented by the drug that would subsequently be discharged. Therefore, the advice would not really place the advisee in a better situation than the one he would end up in anyway. However, this does not suffice to show that

see it as morally wrong to advise a person who is clearly agitated to try to calm himself down, even if he does not wish to receive this advice at the time at which it is given; or that one would be in the wrong to propose to a recovering alcoholic not to have the drink he has ordered in the event that he does not want this advice. Indeed, in both of these cases at least, it seems to us morally desirable to present the agitated person and the recovering alcoholic with such advice even if this is information they do not wish to receive. Many more examples could easily be cited. Furthermore, advice encouraging the advisee to act in ways that are beneficial for other people is at least sometimes also permissible, even if the advisee does not wish to receive it. It is, in our view, absurd to suggest that it would be wrong to advise your rich uncle to give a large portion of his wealth to combat extreme poverty if your uncle did not wish to be given this advice. The more general point is that, if the fact that someone does not wish to receive advice can be said to prohibit us from providing it, then many instances in which it is clearly desirable (or, at least, permissible) to present someone with advice they do not wish to receive would seem to be ruled out.

This is not to say that it can never be morally dubious to present someone with unwanted advice. For instance, if one is constantly bombarded with unwanted advice on how to act (say, every thirty seconds), this is sure to be highly disruptive of, for example, one's ability to direct one's own life. Similarly, if the interventionist advisory device were to constantly advise the offender to engage in countermeasures, this may be highly debilitating in similar ways and surely a strong moral reason to oppose the device's use. Whether some offenders might experience such an extreme stream of advice from the device, however, is ultimately an empirical question. But it seems unlikely to become a widespread practical problem given that the groups of offenders under consideration, i.e., those suffering from uncontrollable episodes of explosive aggression, presumably do not experience such episodes at a frequency that would result in constant behavioural advice from the device.[9]

---

the advice would not be beneficial. For instance, it might be the case that it would be more satisfactory to the advisee if he were to handle the situation himself by following the advice from the device, than if the aggressive outburst was prevented by the drug.

9. This presumption derives some plausibility from a study indicating that, on average, the highest number of episodes experienced during a single year by individuals suffering from intermittent explosive disorder (a disorder characterized by episodes of impulsive explosive aggression) is 27.8 episodes (Kessler et al. 2006).

All in all, based on the above considerations, we believe an objection to employing an interventionist advisory device on the basis of it potentially involving the presentation of advice to the offender against his will should be rejected.

4. Treatment Due to an Act of Omission

Suppose it is true, as has been argued in the previous sections, that the non-consensual collection of neural data and the advice-giving function of the brain device should not lead us to reject its use on some offenders. There might still be reasons related to the interventionist feature of the device, i.e., its ability to discharge countermeasures to prevent an explosive aggressive episode from materializing, that would speak against its use. One such reason relates to whether the intervention of the device would violate the offender's autonomy.

Given the prominence ascribed to the value of personal autonomy in contemporary bioethics and beyond, it is not surprising that a central question regarding the coercive use of neurointerventions on offenders has been whether such use violates an offender's autonomy and, if so, under what conditions (if any) this is morally permissible (see, e.g., Ryberg 2020, chapter 2; Douglas et al. 2013; Caplan 2006). As might be clear, however, this does not seem to be a concern that could plausibly be raised in relation to the interventionist function of the brain device under consideration here. Specifically, the fact that the offender would have the option of preventing the device from discharging countermeasures to stop the aggressive episode from materializing would arguably ensure that his autonomy regarding whether to receive the intervention remains intact.[10] However, it may be objected that the fact that the device in this regards operates as an opt-out system (in which countermeasures are discharged as a result of an act of omission

10. It may be objected that a person who is about to experience an episode of impulsive aggression may not be competent to decide whether to receive treatment. That is, it may be suggested that the impulsive aggression might cloud his decision-making to such an extent so as to make him non-autonomous. However, whether this is indeed a viable concern is ultimately an empirical question regarding whether the device is able to predict the occurrence of the aggressive episode prior to it affecting his decision-making capacities. Furthermore, and perhaps more importantly, while it can surely be ethically dubious to subject a non-autonomous individual to a treatment (e.g., if the said treatment is not in the individual's best interest), the absence of a capacity for autonomy means that doing so cannot plausibly be a violation of his autonomy. Therefore, if an offender would in fact be non-autonomous due to a heightened level of aggression in most (or all) cases in which he must decide whether to receive the forthcoming treatment, the device can hardly be said to violate his autonomy.

on the part of the offender) provides ground for doubting that receiving the intervention from the device is truly the offender's autonomous decision. Specifically, two concerns regarding opt-out systems – often voiced in the debate about implementing such system to increase the availability of organs for transplantation – might be raised against this feature of the interventionist advisory device. Let us consider each in turn.

First, an opt-out system should not be too difficult for persons to opt out of. If it is too difficult (or even practically impossible) for persons to leave an opt-out system, it may plausibly be argued that this system impedes autonomous decision-making. This is surely true, but it does not seem a relevant challenge to mount against a proposal of using advisory interventionist devices on some offenders. More precisely, since the device could allow the offender to reject receiving the proposed treatment with a push of, for example, a button on the device itself or a handheld device, it should hardly be labelled as too difficult to avoid.

Second, and perhaps more importantly, in debates on the ethics of organ procurement it is sometimes argued that an opt-out system relies on the ethically dubious notion of *presumed consent*. That is, it is inferred from a failure to opt out of the system that a person would have given his consent had he been asked to express an explicit view on the matter. One major concern is that assuming consent in this manner overlooks that the person might simply have failed to register his dissent due to, for example, ignorance regarding his registration in the system. Similarly, it could be argued that, when the device discharges its countermeasures, the fact that the offender has not instructed the device not to do so (i.e., he has not opted out) cannot be assumed to mean he would have consented to the treatment had he been asked to express an explicit view on the matter. This cannot be assumed because the offender may simply be ignorant of the fact that the device is going to discharge countermeasures. Should this concern lead us to reject the use of the device under consideration? We believe the answer should be in the negative. First, it is widely believed that it is often morally appropriate to presume consent for treatment from individuals from whom it is difficult or impossible to collect explicit consent if not doing so will result in grievous harm to them. This is, for example, why most of us believe it is usually morally uncontroversial to subject an unconscious victim of a traffic accident to medical treatment without her explicit consent. Similarly, it could plausibly be argued that, if an offender is not in a position to explicitly consent or dissent to the device's treatment due (for example) to ignorance, it may be morally permissible to presume his consent if his explosive aggressive episodes are likely to lead to tremendous harm to him. Therefore, even if it is indeed the case that the device presumes an offender's consent, it is not obvious that this is always morally wrong. Second, it is not

at all clear that the interventionist advisory brain device we have described would in fact presume consent to its treatment. As described above, the brain device would, through a handheld device or the like, inform the offender prior to each instance of discharging its aggression-hampering treatment that the treatment will commence unless he actively keeps it from doing so. Thus, the offender would under normal circumstances seem to be fully informed about the forthcoming treatment and well-positioned actively to decide for himself whether or not he wishes to receive it.[11] Consequently, if and when the device delivers its treatment, it does not seem to have presumed the offender's consent.

In summary, the fact that the countermeasures discharged by the advisory interventionist device would come about due to an act of omission on the part of the offender does not seem to be cause for moral concern.

5. Preventing Explosive Aggression with Stimulation or Drugs

It may, however, be argued that, even if it is true that the interventionist feature of the device under consideration does not violate an offender's autonomy when it discharges its countermeasures, there is another, more basic, problem with this feature related to the means the device employs to prevent aggressive episodes. Specifically, it may be argued that, regardless of whether the treatment violates autonomy, it is simply morally inappropriate to prevent episodes of explosive aggression by means of brain stimulation or aggression-hampering drugs. As Martha Farah has pointed out, using these techniques to reduce aggression instead of more traditional approaches such as anger management classes "renders the effect no less therapeutic. Yet many people's intuitions raise a flag here. And if not here, then at the thought of more permanent interventions such as implanted stimulators or neurosurgery to achieve the same goals" (Farah 2002, 1126). However, while such a means-based argument against employing interventionist advisory devices may have great intuitive appeal, further scrutiny reveals it to face at least two serious challenges.

First, it should be acknowledged that other treatment schemes, such as cognitive behavioural therapy, may turn out to be more effective in preventing explosive aggression in offenders than the proposed device. This is ultimately an empirical question, and one we are currently not in a position to answer. However, should it turn out that

11. As noted above, we will not enter into a more precise discussion of how the device should be designed to be able to deliver the advice to the offender most successfully, but obviously there are many possibilities (e.g., vibration of the handheld device; a spoken message; a particular ringtone; etc.).

the interventionist advisory device is both safe to use and the most effective way of preventing explosive aggression in offenders, it is difficult, in our view, to see why the fact that this effectiveness is ensured by brain stimulation or the discharge of a drug should be considered morally problematic. Second, if the wrongness of employing the advisory interventionist brain device on offenders arises from the wrongness of the means it uses to prevent aggressive episodes, these means must surely also be considered morally wrong to use in other cases as well. This would, however, seem to imply that using brain stimulation techniques or drug-based treatment schemes to, for example, treat individuals suffering from mental health problems should *generally* be considered morally dubious. Surely, few (if any) would accept this view.

However, there is an alternative way that an opponent of using drugs or brain stimulation to prevent explosive aggressive episodes could motivate this view. He or she could argue that what is morally important is not simply that stimulation or medication is being used to prevent explosive aggressive episodes; rather, while these episodes may be socially undesirable, they are non-pathological, and non-pathological conditions should not be treated by means of drugs or brain stimulation. However, while this variation of the objection is surely more plausible than the variation considered above, it still faces at least one crippling challenge. The challenge starts from the observation that there are countless examples where we accept the use of drugs to treat non-pathological states. It is, for example, not usually considered morally problematic to take a sleeping pill to avoid the occasional sleepless night. The same is the case with occasionally taking a pain reliever to treat a headache or a sore knee. But if one insists that it is wrong to treat non-pathological states with drugs, then these and many other similar cases should be taken to involve acts that are wrong to perform. However, surely an account which implies that clearly morally innocuous acts (such as taking a pain reliever to combat the occasional headache) should be morally dubious to engage in is itself highly dubious.

To sum up this section, what seem to us the two most plausible variations of an argument against employing interventionist advisory device turning on the allegedly morally problematic means it uses to prevent episodes of explosive aggression, both seem to face the challenge of becoming overinclusive. It is not clear, at least to us, whether and, if so, how one could specify the objection in a way that avoids this problem.

6. Being Coerced to Wear the Device

As has been argued above, one of the advantages of the proposed interventionist advisory device is that it plausibly does not violate offenders' autonomy when it discharges its countermeasures since it informs offenders of the fact that the treatment is about to commence and leaves them free to reject the said treatment should they wish to do so. It may, however, be argued that there is another way in which the device may be an affront to an offender's autonomy. Specifically, it may be suggested that, even if it is assumed (as we have) that the device would not cause offenders discomfort or other side-effects that may plausibly be debilitating to their ability to exercise their autonomy, requiring an offender to have the device placed on his body against his will is an autonomy violation in its own right.[12] However, while it is surely plausible to hold that individuals should usually be considered the final arbiters concerning what is placed on their bodies, it is not obvious that this shows that interventionist advisory devices should not be used on some offenders.

First, in the context of criminal justice, offenders are often required to place items on their bodies that they may not wish to have placed there, but such requirements are usually not considered morally questionable. Some offenders serving their time outside of prison may, for example, be required to wear an electronic tag (usually placed around their ankle) that monitors their location. And, while perhaps more controversial, some jurisdictions require inmates to wear prison uniforms while incarcerated.[13] It is not clear, at least to us, whether there is a relevant difference between (presumably morally acceptable) practices requiring offenders to wear these objects on their body and requiring them to wear a brain device.

Second, and more generally, the criminal justice system is rife with practices that reduce offenders' autonomy but are nevertheless usually considered morally permissible (or even desirable). For example, it is usually accepted that incarceration, at least in some

12. This objection may plausibly be framed, not in terms of an autonomy violation, but as a violation of offenders' right to self-ownership (see, e.g., Thomson 1990). However, the challenges we offer seems to us to apply regardless of the objection's specific moral foundation.

13. It should, however, be noted that the reason why requiring offenders to wear prison uniforms is most often held to be morally controversial is not that being coerced to wear them violates their autonomy; rather, it is that prison uniforms stigmatize offenders. It is also worth pointing out that, if our proposed brain device could be placed somewhere discreet on the offender (such as behind his ear), it could hardly be said stigmatize offenders in a similar way.

cases, can be a morally appropriate response to wrongdoing, even though it involves restricting the control of offenders over their own life in the form of, *inter alia*, constraints on free movement and association. Arguably, relative to these and other constraints entailed by incarceration, the violation of offenders' autonomy by a device being placed on their body against their will seems, at least in our view, much less severe.

In sum, it is not clear that one can consistently accept the use of incarceration and many other criminal justice practices that impede offenders' autonomy while rejecting the use of an interventionist advisory device on offenders on the basis of its being placed on their bodies against their will.[14]

7. The Impact of the Device on Offenders' Self-Conception

Suppose the interventionist advisory brain device, based on an offender's neural data, has predicted that the offender is about to experience an explosive aggressive episode and has advised the offender that it will discharge countermeasures unless he actively keeps it from doing so and that the offender has allowed the device to commence treatment. Suppose, further, that it is true (as we have argued in previous sections) that none of these steps should raise moral suspicion. There may yet be a reason to be sceptical of employing the device, because the changes to an offender's character induced by the treatment could potentially have an impact on their self-perception[15] – that is, roughly, the experience offenders have of themselves after the intervention from the device. This could be said be to be a relevant concern because some subjects, having received treatment through other brain devices (such as Deep Brain Stimulation), have offered such reports as "I don't feel like myself anymore" after the device was installed (Schüpbach et al. 2006, 1813; see also Baylis 2013, 514). Such post-intervention testimonials have been framed as experiences of loss of authenticity, self-estrangement, or self-alienation (e.g., Kraemer 2013; Gilbert 2018; Pugh, Maslen, and Savulescu 2017). There are, however, several reasons why an objection to the use of interventionist

14. To avoid misunderstandings, it should be underlined that we do not presuppose that the way the criminal justice system currently treats criminals is morally acceptable. In our view there are strong reasons against current trends of mass incarceration. All we are suggesting is that there some cases in which it is acceptable to use incarceration as a punishment.

15. A related worry is that the mere knowledge of having the device on or in one's body may have a negative impact on one's self-conception. Whether this would indeed be the effect of receiving our proposed implant is a question that requires further empirical scrutiny.

advisory brain devices based on the effect they may have on an offender's self-perception does not seem to us to be convincing.

An initial observation worth making is that self-estrangement and the like are not always experienced as negative by the affected person. In a recent study involving patients treated for Parkinson's disease by means of Deep Brain Stimulation, for example, the authors found that most of the patients considered the estrangement induced by the treatment to be restorative in nature (Gilbert et al. 2017). That is, roughly, to these patients, the changes to their character induced by the treatment were seen as restoring elements of their self that had been subdued by the disease (see also Pugh 2020). Thus, even if the impact of our proposed brain device might lead some or all of its recipients to experience self-estrangement, self-alienation, or inauthenticity, it cannot be straightforwardly concluded that this is a cause for moral concern. But even if we, *arguendo*, assume that offenders will generally not welcome changes to how they experience themselves post-intervention, it is not clear that this shows that the brain device should not be employed. At least, one cannot consistently accept the use of imprisonment, which studies have demonstrated to have numerous psychological effects on offenders that may plausibly affect their self-conception (e.g., Haney 2002, 82), while rejecting the use of the proposed device. As already noted, however, most of us believe that incarceration can at least in some cases be a permissible (or even desirable) way of responding to some kinds of wrongdoing.

More importantly, although it is of course ultimately an empirical question, there are reasons to speculate that the treatment delivered by the device under consideration will most likely not lead offenders to experience a loss of authenticity or the like. Much, of course, depends on what exactly causes such experiences to emerge, but the proposed device would at least not seem to give rise to some obvious source for such an experience. For example, the fact that the offender is fully informed about and has full control over the device when it delivers its treatment seems to make it unlikely that he will be troubled by the uncertainty of not knowing when the device is affecting his behaviour. These same features of the device may also ensure that the changes to his character induced by the device are not experienced as the product of an alien intrusion.

Furthermore, some subjects in fact report feeling *more* like themselves after receiving certain forms of neurotechnological treatment. The perhaps most cited example of this is the use of the antidepressant Prozac (see, e.g., Kraemer 2013, 486). It seems plausible, in our view, that a treatment meant to reduce impulsive aggression, like our proposed brain device, might well have a similar effect. After all, an offender's behaviour during episodes of impulsive aggression when he is not fully in control of how he acts is presumably not

experienced by the offender as reflecting who he really is. If this is true, then a device like the one proposed to give offenders more control over their behaviour may plausibly leave them feeling more in line with who they perceive themselves to be (see also Ryberg 2012, 233).

To conclude this section, we have offered reasons to doubt that the potential impact of the treatment from an interventionists advisory brain device on offenders' self-conception would be experienced negatively by the affected offenders, supposing that such effects are indeed likely to emerge. It has also been argued that there is, in fact, reason to doubt that the device would have such effects – indeed, the device may plausibly aid offenders in behaving in ways that are more in line with how they perceive themselves.

8. Conclusion

Some novel brain devices currently being investigated for use in a clinical setting are able to predict and advise patients about the emergence of specific neural events, thus making patients capable of engaging in the appropriate countermeasures. Similar devices are able to automatically adjust the timing, intensity, and location of treatment to counteract unwanted neural events. Inspired by these developments, we have considered some important ethical questions related to employing a hypothetical brain device combining such advisory and interventionist features to reduce recidivism among offenders suffering from severe problems with impulsive aggression. This device would have three features: (1) it would monitor an offender's brain data to predict upcoming aggressive episodes; (2) it would offer the offender behavioural advice; and (3) it would, unless the offender actively kept it from doing so, administer treatment by discharging measures to ensure that the aggressive episode does not materialize. There are *prima facie* plausible moral reasons to oppose each of these features, but we have suggested that, on closer scrutiny, none of the reasons considered convincingly rules out mandating advisory interventionist brain devices to the specified group of offenders. This conclusion does not, however, suffice to show that such devices should indeed be used. There are, for example, important ethical questions regarding the implementation of the device that we have not addressed. Much of this discussion would seem to hinge on exactly how invasive a procedure would be needed to implement the interventionist advisory device. Surely, if it is only possible to implement such a device by means of invasive brain surgery, this would provide a very strong reason to oppose its use. If, on the other hand, technological

developments mean that the device could be placed on the exterior of an offender's skull (e.g., behind an ear) and assert its effect through a needle placed just below his skin, arguments against mandating the device relating to its implementation seem less appealing. We have argued that the features of the device outlined above should not be cause for moral concern; therefore, *if* ethically unproblematic means of implementation are indeed developed, it is not obvious to us what principled reasons may be offered to oppose the use of interventionist advisory brain devices on some offenders.

References

Baylis, Françoise. 2013. "'I Am Who I Am': On the Perceived Threats to Personal Identity from Deep Brain Stimulation." *Neuroethics* 6 (3): 513–26. https://doi.org/10.1007/s12152-011-9137-1.

Birks, David, and Thomas Douglas. 2018. *Treatment for Crime: Philosophical Essays on Neurointerventions in Criminal Justice*. Oxford University Press.

Bublitz, Jan Christoph, and Reinhard Merkel. 2014. "Crimes Against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination." *Criminal Law and Philosophy* 8 (1): 51–77. https://doi.org/10.1007/s11572-012-9172-y.

Caplan, Arthur L. 2006. "Ethical Issues Surrounding Forced, Mandated, or Coerced Treatment." *Journal of Substance Abuse Treatment* 31: 117–20.

Cook, Mark J., Terence J. O'Brien, Samuel F. Berkovic, Michael Murphy, Andrew Morokoff, Gavin Fabinyi, Wendyl D'Souza, et al. 2013. "Prediction of Seizure Likelihood with a Long-Term, Implanted Seizure Advisory System in Patients with Drug-Resistant Epilepsy: A First-in-Man Study." *The Lancet Neurology* 12 (6): 563–71. https://doi.org/10.1016/S1474-4422(13)70075-9.

Douglas, Thomas. 2014. "Criminal Rehabilitation Through Medical Intervention: Moral Liability and the Right to Bodily Integrity." *The Journal of Ethics* 18 (2): 101–22. https://doi.org/10.1007/s10892-014-9161-6.

Douglas, Thomas, Pieter Bonte, Farah Focquaert, Katrien Devolder, and Sigrid Sterckx. 2013. "Coercion, Incarceration, and Chemical Castration: An Argument From Autonomy." *Journal of Bioethical Inquiry* 10 (3): 393–405. https://doi.org/10.1007/s11673-013-9465-4.

Farah, Martha J. 2002. "Emerging Ethical Issues in Neuroscience." *Nature Neuroscience* 5 (11).

Gilbert, Frederic. 2015. "A Threat to Autonomy? The Intrusion of Predictive Brain Implants." *AJOB Neuroscience* 6 (4): 4–11. https://doi.org/10.1080/21507740.2015.1076087.

Gilbert, Frederic. 2018. "Deep Brain Stimulation: Inducing Self-Estrangement." *Neuroethics* 11 (2): 157–65. https://doi.org/10.1007/s12152-017-9334-7.

Gilbert, Frederic, Eliza Goddard, John Noel M. Viaña, Adrian Carter, and Malcolm Horne. 2017. "I Miss Being Me: Phenomenological Effects of Deep Brain Stimulation." *AJOB Neuroscience* 8 (2): 96–109. https://doi.org/10.1080/21507740.2017.1320319.

Glannon, W., and C. Ineichen. 2016. "Philosophical Aspects of Closed-Loop Neuroscience." In *Closed Loop Neuroscience*, edited by Ahmed El Hady. Elsevier/Academic Press.

Haney, Craig. 2002. "The Psychological Impact of Incarceration: Implications for Post-Prison Adjustment." *U.S. Department of Health & Human Services*. http://img2.timg.co.il/CommunaFiles/19852476.pdf.

Holmen, Sebastian Jon. 2020. "Respect, Punishment and Mandatory Neurointerventions." *Neuroethics* (May). https://doi.org/10.1007/s12152-020-09434-8.

Ienca, Marcello, and Roberto Andorno. 2017. "Towards New Human Rights in the Age of Neuroscience and Neurotechnology." *Life Sciences, Society and Policy*. https://doi.org/10.1186/s40504-017-0050-1.

Kessler, Ronald C., Emil F. Coccaro, Maurizio Fava, Savina Jaeger, Robert Jin, and Ellen Walters. 2006. "The Prevalence and Correlates of DSM-IV Intermittent Explosive Disorder in the National Comorbidity Survey Replicatication." *Archives of General Psychiaty* 63 (June): 718–32.

Kirchmair, Lando. 2019. "Objections to Coercive Neurocorrectives for Criminal Offenders – Why Offenders' Human Rights Should Fundamentally Come First." *Criminal Justice Ethics* 38 (1): 19–40. https://doi.org/10.1080/0731129X.2019.1586216.

Kraemer, Felicitas. 2013. "Me, Myself and My Brain Implant: Deep Brain Stimulation Raises Questions of Personal Authenticity and Alienation." *Neuroethics* 6 (3): 483–97. https://doi.org/10.1007/s12152-011-9115-7.

Lavazza, Andrea. 2018. "Freedom of Thought and Mental Integrity: The Moral Requirements for Any Neural Prosthesis." *Frontiers in Neuroscience*. https://doi.org/10.3389/fnins.2018.00082.

Ligthart, Sjors, Tijs Kooijmans, Thomas Douglas, and Gerben Meynen. Forthcoming. "Closed-Loop Brain Devices in Offender Rehabilitation: Autonomy, Human Rights,and Accountability." *Cambridge Quarterly of Healthcare Ethics*.

Malekmohammadi, Mahsa, Jeffrey Herron, Anca Velisar, Zack Blumenfeld, Megan H. Trager, Howard Jay Chizeck, and Helen Brontë-Stewart. 2016. "Kinematic Adaptive Deep Brain Stimulation for Resting Tremor in Parkinson's Disease." *Movement Disorders* 31 (3): 426–28. https://doi.org/10.1002/mds.26482.

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. https://doi.org/10.1136/medethics-2016-103492.

Pugh, Jonathan. 2020. "Clarifying the Normative Significance of 'Personality Changes' Following Deep Brain Stimulation." *Science and Engineering Ethics* no. 0123456789. https://doi.org/10.1007/s11948-020-00207-3.

Pugh, Jonathan, Hannah Maslen, and Julian Savulescu. 2017. "Deep Brain Stimulation, Authenticity and Value." *Cambridge Quarterly of Healthcare Ethics* 26 (4): 640–57. https://doi.org/10.1017/S0963180117000147.

Pycroft, Laurie, Sandra G. Boccard, Sarah L.F. Owen, John F. Stein, James J. Fitzgerald, Alexander L. Green, and Tipu Z. Aziz. 2016. "Brainjacking: Implant Security Issues in Invasive Neuromodulation." *World Neurosurgery* 92: 454–62.

Ryberg, Jesper. 2012. "Punishment, Pharmacological Treatment, and Early Release." *International Journal of Applied Philosophy* 26 (2): 231–44. https://doi.org/10.5840/ijap201226217.

Ryberg, Jesper. 2015. "Predictive Brain Devices, Therapeutic Activation Systems, and Aggression." *Ajob Neuroscience* 6 (4): 36–38. https://doi.org/10.1080/21507740.2015.1094548.

Ryberg, Jesper. 2017. "Neuroscience, Mind Reading and Mental Privacy." *Res Publica* 23 (2): 197–211. https://doi.org/10.1007/s11158-016-9343-0.

Ryberg, Jesper. 2020. *Neurointerventions, Crime, and Punishment: Ethical Considerations*. New York: Oxford University Press.

Schüpbach, M., M. Gargiulo, M.L. Welter, L. Mallet, C. Béhar, J.L. Houeto, D. Maltête, V. Mesnage, and Y. Agid. 2006. "Neurosurgery in Parkinson Disease: A Distressed Mind in a Repaired Body?" *Neurology* 66 (12): 1811–16. https://doi.org/10.1212/01.wnl.0000261103.50365.5b.

Shaw, Elizabeth. 2014. "Direct Brain Interventions and Responsibility Enhancement." *Criminal Law and Philosophy* 8 (1): 1–20. https://doi.org/10.1007/s11572-012-9152-2.

Thomson, Judith Jarvis. 1990. *The Realm of Rights*. Cambridge, MA: Harvard University Press.

Journal of Cognition and Neuroethics

Exploring Moral Bio-enhancement through Psilocybin-Facilitated Prosocial Effects

Victor Lange
University of Copenhagen

Sidsel Marie
University of Copenhagen

Acknowledgements
We would like to thank you the members of CEEC, at the Philosophy Department of the University of Copenhagen, for an interesting discussion and comments on an earlier draft of this paper.

Biographies
Victor Lange is a PhD-fellow at the Section for Philosophy and the Centre for Neuroscience at the University of Copenhagen. His PhD-project investigates person-level metacognitive control, as it is studied in clinical and performance psychology, with a perspective on the philosophical dimensions of agency, attention, and introspection. He has previously done research in bioethics (also at the University of Copenhagen). In addition, he currently works as a meditation teacher and editor at the platform Regnfang.

Sidsel Marie is a master-student at the Department for Anthropology at the University of Copenhagen. Her work investigates integration of psychedelic experiences into everyday-life among Danish psychedelic users, with a particular focus on the involved social dynamics. Sidsel Marie is further editor and yoga and meditation teacher at the platform Regnfang.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). September, 2021. Volume 8, Issue 1.

Citation
Lange, Victor, and Sidsel Marie. 2021. "Exploring Moral Bio-enhancement through Psilocybin-Facilitated Prosocial Effects." *Journal of Cognition and Neuroethics* 8 (1): 23–64.

Exploring Moral Bio-enhancement through Psilocybin-Facilitated Prosocial Effects

Victor Lange and Sidsel Marie

Abstract

The idea of moral bio-enhancement has received considerable philosophical attention in the last 10 years. Yet, it has been extremely difficult to come up with plausible and feasible procedures for how to perform such enhancement. The purpose of this paper is to explore whether the psychedelic compound psilocybin, due to its prosocial effects, can be used for moral bio-enhancement. The first part of the paper is conceptual. This part investigates the term 'prosociality', relates it to philosophical discussions of moral bio-enhancement, and presents a set of necessary conditions for when increases in prosociality can count as moral enhancement. The second part of the paper reviews the empirical literature on the prosocial effects of psilocybin. This part proposes that the prosocial effects of psilocybin likely satisfy the above-mentioned set of six necessary conditions. The paper hereby proposes that we have reason to be tentatively and moderately optimistic about using psilocybin for moral bio-enhancement and that this use of psilocybin is worth future research attention. Nonetheless, the paper ends by stressing that both further philosophical and empirical research is crucial for making stronger conclusions on this matter. The last section of the paper suggests a set of outstanding research questions that should be targeted in such future research.

Keywords

Moral Bio-enhancement, Prosociality, Psilocybin, Psychedelics

1. Introduction

The term 'moral bio-enhancement' refers to the process in which an agent becomes a morally better agent with the assistance of some bio-medical or bio-technological entity (Douglas 2008; Earp 2018). This paper concerns two specialized discussions within the current moral bio-enhancement literature.

First, the paper concerns the discussion on whether moral bio-enhancement could be carried out by making agents more prosocial (Section 2.1 clarifies the term 'prosociality'). The paper offers a new approach to this discussion. By reviewing and structuring the relevant philosophical literature, the paper formulates a set of six necessary conditions that states the boundaries for when an increase in prosociality can count as morally enhancing an agent. This set of conditions is intended to represent the common worries of philosophers on performing moral bio-enhancement through increases in prosociality.

We do not argue for the correctness of this set but formulate it to represent positions held in the philosophical literature.

Second, the paper concerns whether psilocybin (the main psycho-active ingredient in magic mushrooms and psychedelic truffles) could be an appropriate mean for moral bio-enhancement. Other philosophers have already discussed the potential of psilocybin for moral bio-enhancement (Earp 2018; Tennison 2012; see also Haidt 2012). Nonetheless, this paper offers an in-depth analysis of the prosocial effects of psilocybin. The paper does so by applying the framework of the above-mentioned set of six necessary conditions. Such a detailed examination of psilocybin for moral bio-enhancement has not been presented before.

While the main purpose of the paper is to explore these matters, the main claim of the paper is that we have reasons to be tentatively and moderately optimistic about psilocybin for moral bio-enhancement. The paper makes this claim by stressing that the prosocial effects of psilocybin appear to satisfy the set of six necessary conditions mentioned above. Hereby, the paper stresses that the use of psilocybin for moral bio-enhancement is a topic worthy of future research attention.

The structure of the paper is as follows. Section 2 is of a more conceptual nature and offers an introduction to the terms of 'prosociality' and 'moral bio-enhancement'. Most importantly, the section presents the above-mentioned set of six necessary conditions. Section 3 is of a more empirical nature and offers an analysis of the scientific literature on the prosocial effects of psilocybin. This section relates these prosocial effects to the mentioned set of necessary conditions. At last, Section 4 suggests some important research questions for future research on psilocybin for moral bio-enhancement.

To avoid misunderstandings, it is fruitful to stress that the present paper has an explorative nature: its purpose is to explore whether the idea of performing moral bio-enhancement through psilocybin has any initial plausibility. This means that the paper leaves some central philosophical aspects and discussions of moral bio-enhancement and prosociality untouched and unsettled. Further, the paper remains silent about some deeper philosophical questions on moral psychology and moral disagreement. This might be a disappointment to some philosophical readers. This prioritization is fully intended nonetheless. As with any paper, this paper operates under a limited scope. We have given priority to provide a more applied and practical perspective on psilocybin, prosociality, and moral bio-enhancement—hereby, leaving important meta-ethical, normative, and moral psychological topics under-discussed. The literature on moral bio-enhancement is scarce with real feasible and practical proposals on bio-medical substances

or neurotechnologies for moral bio-enhancement. This paper, first and foremost, aims to meet this lack of practical proposals.

2. Prosociality and Moral Bio-enhancement

This section provides an initial and rough clarification of the term 'prosociality'. Importantly, the reader should notice that the purpose of this section is to review and structure the relevant scientific and philosophical literature. The following clarifications are not intended to constitute an independent argument on how to properly understand prosociality and its relation to moral bio-enhancement. Instead, the section aims to adequately represent how the term of 'prosociality' is used across research disciplines. The section furthermore aims to structure the main worries that philosophers have for performing moral bio-enhancement through prosociality. The purpose of this section is hereby primarily to represent the views of different authors, not to evaluate it. We provide this review and structuring of the relevant literatures to build a broad conceptual framework through which Section 3 can evaluate whether psilocybin meets the commonly held worries of the philosophical and bio-ethical community.

2.1 Characterising Prosociality

The term 'prosociality' is used in various research disciplines to characterise individuals' attitudes and behaviour. The most common way to use the term could be outlined as follows:

> *Prosocial attitude*. An agent, A, holds a prosocial attitude towards an individual or a group, G, to the degree that she is motivated to benefiting G.
>
> *Prosocial behaviour*. An agent, A, behaves prosocially towards an individual or group, G, if A's behaviour is driven by a prosocial attitude towards G.[1]

1. In both definitions it is assumed that A and G are non-identical, even though (supposing G is a group) the definitions allow A to be a group-member of G.

Note five things about the above two outlines. First, the definitions are morally neutral and very broad. They make no requirements on specific moral features of the relevant attitudes or behaviour (such as whether the underlying motivation of the attitude or behaviour is genuinely other-concerned). This morally neutral and broad character of the definitions is important to notice, and we shall discuss it further below. Second, the definitions take prosocial attitudes and behaviour to be interpersonal. They are attitudes that are essentially characterised by how an individual approaches or relates to other individuals or groups. Third, according to the above outline, prosocial attitudes come in degrees since the motivational state to benefit some individual or group can come with varying strength. Four, it is important to stress that the expression 'to benefit' should be understood rather broadly in the definitions. Here, 'to benefit' refers to both more economical or material kinds of benefiting (i.e., in which an agent is motivated to distribute resources to benefit some individual or group) and more emotional forms in which a person is motivated to benefit an individual or group through emotional support and relational investment (such as in providing comfort and reconciliation. See: de Waal & van Roosmalen 1979). Of course, both material and more emotional forms of benefitting take various forms, depending upon what the relation that obtains between the agents or groups involved (Clark & Mills 1993, 2012; Clark & Taraban 1991; Earp et al. 2020). We touch further upon this in Section 3 below. Five, it is clear by the above outlines that prosocial behaviour presupposes prosocial attitudes (Twenge et al. 2007).

These two outlines, and the elaboration of them, explicate how the term 'prosociality' is characterised and used in scientific research—involving psychology (Jensen et al. 2014; Zaki & Mitchell 2013; Twenge et al. 2007, 56), evolutionary biology and primatology (de Waal 2006, 2008; Jensen 2016). Further, research on psilocybin (and other psychedelics) also applies this usage (Milliére et al. 2018; Griffiths et al. 2018). More popular characterisations, such as Wikipedia articles, also account for prosociality as done by the two accounts (Wikipedia 2021).

Researchers talk about certain attitudes (or emotion-like states) and kinds of behaviour as being paradigmatic examples of prosociality. Paradigmatic 'prosocial' attitudes and emotion-like states would be empathy, sympathy, trust, a desire for being close to others, and a felt connection to others (de Waal 2006, 2008; Trautwein et al. 2014; Batson 1991). These emotion-like states are seen as reliable indicators of present prosocial attitudes in individuals, or as reliable producers of prosocial attitudes. Examples of paradigmatic types of prosocial behaviour would be different forms of altruism, sharing, and helping (Jensen et al. 2014). These types of behaviour are seen as reliable indicators of an individual holding a prosocial attitude towards the relevant

other individuals or groups. Researchers think that although numerous factors (such as situational issues, lack of will-power in the relevant individual, or other overriding motivational states) might cancel out prosocial behaviour, increasing the degree to which an individual holds prosocial attitudes, toward other individuals or groups, generally makes the individual more prone to behave prosocially towards the relevant individual or group (Zaki & Mitchell 2013; Twenge et al. 2007).

Prosociality is moreover often characterised as a kind of attitude and behaviour opposed to so-called antisocial attitudes and behaviour (Basurto et al. 2016). Antisocial attitudes concern the motivation to disadvantage someone or the indifference to other individuals' well-being if one can gain benefits from exploiting them. Paradigmatic examples of antisocial attitudes of emotion-like states are different instances of aggression, and examples of antisocial behaviour would be violence and unwillingness to cooperate (Hilton et al. 2018). Psychiatric conditions such as psychopathy are associated with antisociality (Neumann et al. 2015). With this in mind, prosociality and antisociality are often seen as two ends on a spectrum, meaning that increase in one of them leads to decrease in the other.

Some readers might find the above characterisation of prosociality to be problematic, however. That is, one might object to the two above outlines by stating that the characterisations are too liberal. One way to explicate this objection would be by presenting the following considerations. Consider the below cases in which a person, A, holds a prosocial attitude and might behave prosocially to an individual or group, G, in the following way:

(a) A is motivated to benefit G but refrains from action (for example due to lack of will-power).

(b) A is motivated to benefit G, but ends up disadvantaging S (for example due to lack of knowledge about how to benefit G).

(c) A is motivated to benefit G and behaves accordingly, but A's attitude and behaviour is strongly socially expected (for example a parent is motivated to benefit her child by providing her food and behaves accordingly).

(d) A is in a given situation motivated to benefit G and behaves accordingly. However, the motivation of A to benefit G is only instrumental in the sense that A ultimately seeks to benefit herself

(for example A might help G, but only with the intention of gaining a good reputation in the group or committing G to help her later on).

(e) A is motivated to benefit G, though this motivation is of a 'tribal nature'[2], meaning that the members of G are only in-group members to A, and an antisocial attitude towards some out-group members, G*, is associated with A's increased prosocial attitude to G (for example A might intend to benefit G by disadvantaging some out-group members, G*, by acts of violence, punishment, or some other kind of hostility).[3]

The cases (a)-(e) would all count as instances of prosocial attitudes and behaviour under the definitions given above. This might lead the reader to believe that (a)-(e) constitutes a series of counter-examples, since the reader might think that it would not make sense to characterise these as instances as instances of prosociality.

The purpose of this paper is not to defend how the term 'prosociality' is used in contemporary research practice. We can only stress that the two definitions given above correspond fully with how researchers use the term. In accordance with this, researchers would not take case (a)-(e) to show that the common definitions of prosociality are mistaken. Instead, researchers would most probably take these cases to stress two important aspects concerning the term of 'prosociality' (two aspects we have already flagged above). First, they show that the characterisation of 'prosociality' is morally neutral in the sense that it covers instances of attitudes and behaviours of varying moral character (ranging from morally admirable to problematic character). This basic morally neutral nature of prosociality makes its link to morality non-trivial. As we shall see in the following sections, suggesting moral bio-enhancement to work through some increase in prosociality requires that the increase in prosociality meets certain further requirements. Second, and related, (a)-(e) simply show that the characterisation is minimal in the sense that it covers many different kinds of attitudes and behaviour without, in a fine-grained way, differentiating further between them. These implications correspond fully with how 'prosociality' is discussed and used in the scientific literature. Here, 'prosociality' is exactly

2. Such tribal prosociality is closely linked to the term 'parochial altruism'. See Choi & Bowles (2007).

3. For relevant studies of this 'tribal prosociality', see Van Kleef et al. (2012) and De Dreu et al. (2010).

seen as depicting a very large set of different social attitudes and types of behaviour with varying moral dimensions (Penner et al. 2005; Caprara et al. 2012). Prosociality is simply used as a broad term that depicts the fundamental type of social attitude and behaviour of being motivated to benefits others (an attitude and behaviour that can take multiple forms and have various relations to self-interest, reciprocity, and selfless other-concern).

Some readers might still think that such a morally neutral and broad use of the term 'prosociality' seems strange. These readers might think that prosociality is a morally laden term in the sense that acting prosocially is, other things being equal, a kind of morally positive behaviour and always self-less in some minimal sense. Our purpose is not to disqualify such a morally laden use of 'prosociality'. Yet, as mentioned above, this use is not the dominant one in scientific research. Across different research disciplines (such as primatology, evolutionary biology, social psychology, etc.), prosociality is a very broad term that depicts behaviour proximally caused by an attitude to benefit another individual or group—without implying anything on further morally relevant features of such attitudes and behaviour. The idea behind operating with such a broad term as 'prosociality' is basically to understand the multiple causes and multi-dimensional nature of the attitude and behaviour of proximally intending to benefit others.

2.2 Moral Bio-enhancement through Prosociality

In the following sections, we shall by the expression 'moral bio-enhancement' mean the following (see also Earp et al. 2017, 168):

> Any change in a moral agent, A, effected or facilitated in some significant way by the application of a biotechnology (e.g., bio-medical substances or neurotechnologies), that results, or is reasonably expected to result, in A being a morally better agent.

As it simply stands, this account is obviously under-specified in one important aspect: namely, the aspects of what it means that an individual becomes a morally better agent (or what it basically means to be a morally good agent).[4] There is deep philosophical

4. Some readers might think that this account is unfair to consequentialistic perspectives on moral bio-enhancement. Under such perspectives moral bio-enhancement is merely oriented around the consequences of action, meaning that moral bio-enhancement is basically the processes of increasing the degree or probability of agents to act such that their actions have morally better consequences. Yet, notice that we do not propose this account to be *the* correct understanding of moral bio-enhancement generally.

controversy around what fundamentally characterises a morally good agent. This was already a controversial topic among ancient Greek philosophers, and the disagreement still pertains between major moral theories such as different versions of consequentialism, Kantianism, and virtue ethics (Homiak 2019; see also Miller 2020). Further, religious moral doctrines, such as that of Catholic ethics, provide characterisations of the morally good person in many ways distinct from those of philosophy (Hare 2019). Given this disagreement, some authors think that the idea of moral bio-enhancement is a dead-end: how can we morally enhance individuals when there is no uncontroversial conception of what this means (Beck 2015)?

Though, philosophers proposing procedures of moral bio-enhancement usually try to tackle this profound disagreement by applying a 'convergence approach'. By this approach, authors propose to increase a given trait or capacity which a majority of the plausible moral theories or systems converge upon as being a morally attractive trait or capacity to increase (Shook 2012; Raus et al. 2014; Ahlskog 2017). For example, authors have proposed to diminish a person's racial prejudiced beliefs (Douglas 2008), increase cognitive capacities and abilities of reasoning (Earp et al. 2017, 170; Shaefer 2015), or strengthen will-power (Shook 2012, 5-6), since multiple moral theories would converge upon viewing such enhancements as morally attractive. As is the topic of this paper, authors have also proposed to perform moral bio-enhancement by making individuals more prosocial (Persson & Savulescu 2012; Earp et al. 2017, 170; Shaefer 2015). The fundamental idea is here that generally increasing the motivation of agents to benefit other individuals or groups likely make these agents morally better.[5] Authors proposing this does not claim any strict or necessary link between prosociality and the moral goodness or character of a person. Instead, they seem to suggest that there is a reliable positive association such that increases in prosociality generally make agents morally better.

We adopt it because it is widely used in the philosophical literature. Further, the account allows for a consequentialist to further specify the moral goodness of an agent purely as a matter of the consequences of her actions.

5. Notice that the link between certain kinds of prosociality and moral enhancement is probabilistic in the above outline (by the qualification 'likely'). This does seem to be an appropriate outline: we know of no authors claiming there to be any strict or necessary link between prosociality and the moral goodness of a person, instead the idea seems to be that there is a reliable positive association.

In Section 2.3 below we discuss the many problems of this idea, but let us initially sketch how one could motivate it.[6] We do not claim that this following motivation constitutes a strong and well-developed argument for this idea of performing moral bio-enhancement through increases in prosociality. The purpose is instead to get some sense of how prosociality could overall be seen as convergence point between different moral systems.

One could stress that concerning normative philosophical moral theories, we could reasonably expect that both different versions of consequentialism (Cummiskey 1989) and virtue ethics (see for example many essays in Walker & Ivanhoe 2007) would find the intuition broadly reasonable (since increasing certain prosocial attitudes might lead to promotion of the general welfare or to instantiating virtues). With regards to more descriptive theories of moral psychology, as Ahlskog also points to (2017, 364), plausible conceptions of the evolution of morality stress that morality essentially involves that individuals take each other's well-being into account and are motivated to benefit it (Boehm 2012, 49; de Waal 2006, 3; Kitcher 2011; Joyce 2007; Jensen et al. 2014). The rationale would be that since our capability for moral cognition is often driven by prosociality (at least to some degree), enhancing this motivation might enhance our moral cognition. Moreover, several religious or spiritual moral theories or doctrines, such as Christian ethics (Gill 2012) and Buddhist ethics (Mosig 1989, 27), often emphasise love and compassion as core moral virtues. One might suggest that these systems would probably also find the intuition intially plausible. Many folk conceptions of morality further stress that the morally good person is characterised by her motivation to benefit other people.[7] With these examples in mind, the idea that moral enhancement could be performed by increases in prosociality does seem to have some broad inter-theoretical and

6. For example, some readers might object that the simple intuition, that increases in prosociality can count as moral enhancement, is deeply implausible merely on the ground that it does not make any requirements on the identity of the individual or group targeted by the increased prosociality. Such readers might point to the example that helping a group in executing unjust violence cannot be morally desirable. This issue will be further discussed in Section 2.3 and 2.4 (see especially the discussion of condition (iii) and (iv)), and especially in Section 3.2. However, it is once again important to stress that we do not outline this intuition simply to defend it. Instead, we outline it to represent a basic idea held by some authors in the philosophical literature. As the following sections show, there are several problems connected to this intuition.

7. See for example how Strohminger and Nichols conceptualise moral character in their seminal study (2014). See also the cross-cultural meta-study on morality by Curry et al. (2019), who argue that cooperation, reciprocity, and interpersonal helping behaviour are considered morally good across cultures.

convergence-based motivation. Examining the route of moral bio-enhancement through prosocial effects appears to have some appeal according to various moral systems—at least from a general perspective.

There are several ways in which one could object to the above idea that moral enhancement can be carried out by increasing prosociality and the motivation for it. Consider the following three objections or worries.

(1) One could simply object that the idea fundamentally misunderstands the nature of morality, since morally right actions and being a morally good person have nothing to do with prosociality. For example, an ethical egoist might very well have much reluctance in accepting the conditional (Regis 1980). Further, a Kantian might think of it as mistaking the nature of morality, since moral actions are not driven by prosocial attitudes but by duty to the moral law and in response to the categorical imperative (Johnson & Cureton 2019). If the reader is sympathetic to the first kind of objection, and does not think that there are some interesting positive relations between prosociality and the moral goodness of an agent, then the following sections will be of limited interest. The plausibility of the further analysis of this paper is conditional upon accepting some positive link between prosociality and moral character.

(2) One might alternatively reply to the above motivation that it is way to abstract and under-specific. One might further state that this would mean that further examination would show that the strength of the above proposed convergence is very weak or perhaps non-existent. That is, critical readers might stress that even though different moral systems (e.g., Christian ethics, Buddhist ethics, virtue ethics, utilitarianism) might all suggest certain ways of prosociality to be morally attractive, these systems diverge in their further specification of what kinds of prosocial attitudes and behaviours that more specifically count as morally attractive. The relevant moral theories and systems likely come to disagree heavily in the light of such specification (e.g., utilitarianism most probably disagree with Christian ethics on the dimensions of prosociality involved in abortion). Hence, the convergence, or 'overlapping consensus', by different moral systems on prosociality is not substantial enough to ground the idea of performing moral bio-enhancement through prosocial effects.[8]

This objection points to a fundamental issue: it questions the theoretical support for considering prosociality at all in relation to moral bio-enhancement. We do not aim to

8. We are grateful to an anonymous reviewer for stressing this issue. For further inspiration and a radical example on how differently prosociality can be morally specified by various moral systems, see Rai & Fiske (2012).

settle this issue (it is clearly a demanding task). Yet, we think the issue can be approached in the following way.

First, the general idea that plausible moral theories and systems can converge has already been discussed in the philosophical literature. Such convergence can take various forms in relation to what the point of convergence is. Some philosophers have proposed that major moral theories converge upon higher-level criteria for the moral right action (Parfit 2011). Other philosophers have proposed that major moral theories converge upon mid-level principles that can guide and justify moral decision-making (Beauchamp & Childress 2001). To be clear, when we suggest that moral systems and theories converge upon prosociality as an appropriate target of moral bio-enhancement, we do not mean that these theories and systems take prosociality as being the fundamental criteria for morally right actions. Instead, we mean that these theories and systems take prosociality as standing in some reliable relationship to the morally right action (although, each system or theory will provide independent explanations of how this relationship more exactly is to be understood). The convergence can hereby be understood as an overlapping consensus on prosociality as a practical mid-level approach to enhancing the moral psychologies of individuals.

Second, we take this 'mid-level convergence' on prosociality to be plausible for the following reasons. To begin with, we think that it is reasonable to overall propose that plausible moral theories and systems all agree that selfishness (implying a general lack of motivation to benefit others) is a general and serious moral problem in the sense that it often hinders individuals in performing the morally right action in real life scenarios. This means that diminishing such selfishness and increasing other-concern is overall a strategy that makes initial sense. Further, even though the moral theories and systems would prioritize different kinds of increases in prosociality and potentially disagree on what specific kinds of prosociality that would count as morally good, we expect that they indeed would agree in a sufficiently large number of cases to ground prosociality as mid-level approach to moral bio-enhancement.

At last, the considerations given in Section 3 will also tackle this above worry on the strength of convergence and overlapping consensus. In Section 3, we outline a set of six necessary conditions stating what character an increase in prosociality of an agent must have, for it to count as moral enhancement of that agent. This set of conditions is based upon a review of the relevant philosophical literature. We think that plausible moral theories and systems will largely agree that if an increase in prosociality satisfies these conditions, then it will very likely count as moral enhancement. This means that even if we are wrong in suggesting that there is broadly an overlapping consensus on

prosociality as a mid-level approach to moral bio-enhancement, it does seem reasonable to suggest that there is an overlapping consensus on particular increases in prosociality that satisfy these six conditions. We think that this consideration, together with the above two, does provide some considerable inter-theoretical motivation to explore moral bio-enhancement through prosocial effects.

The above considerations are clearly speculative and might be unsatisfying to the reader. It is important to re-state the main ambitions of this paper. The main claim of this paper is not that plausible moral theories and systems converge upon prosociality as practical mid-level approach to moral bio-enhancement. Instead, the main claim of the paper is that the prosocial effects of psilocybin appear to satisfy the common philosophical worries against performing such enhancement through increases in prosociality. What we have done above is simply to offer a rough motivation for prosociality as a target for moral bio-enhancement. We do not think that this constitutes any strong conclusions on the discussion on prosociality and moral bio-enhancement, yet we do think that it suggest that the idea of moral bio-enhancement through prosociality deserves further philosophical theorising and that is what we offer below.

(3) At last, one could alternatively present a third objection by stressing that the idea of moral bio-enhancement through prosociality is implausible because the link between prosociality and being a morally good person is much more nuanced and non-trivial than described above. Making this objection one might agree that there is indeed a link between prosociality and morality (contrary to what objection (1) denies). Further, one might also agree that there is a relatively robust convergence from multiple plausible moral theories and systems on prosociality as a practical mid-level approach to moral bio-enhancement (contrary to what objection (2) suggests). Objection (3) denies the idea that increasing prosociality likely results in moral enhancement on the ground that multiple other factors profoundly influence the moral character of such increase. In other words, this objection states that if an individual is to become morally better by an increased motivation to benefit some other individuals or groups, this motivation must satisfy a line of other conditions. It is exactly this objection (3) that the following sections discusses. We elaborate upon it below by reviewing the relevant philosophical literature and outline how different philosophers have formulated this objection.

2.3 Elaboration of Objection (3)

The elaboration of objection (3) has taken multiple forms in the philosophical literature. We suggest dividing these into the four following categories (these relate to the cases (a)-(e) discussed in Section 2.1). Once again, it is important to stress that the purpose of outlining these worries is not to argue that these elaboration or worries are in fact all correct—we do *not* aim to provide such an evaluation of them. Instead, the below outline should be seen as a review and structuring of prominent views in the philosophical literature on the problems on performing moral bio-enhancement through increases in prosociality. This is important to stress to understand the purpose of the following sections. Further, it also emphases that the reader, depending upon her general moral outlook, will probably find some of these worries irrelevant. However, since the below serves the purpose of a review, this will not constitute an argument against the present paper, but against those philosophers advocating these worries.

Philosophers have generally opposed the idea of performing moral bio-enhancement through prosocial effects in three ways. We propose to outline these as follows.

Prosociality as non-exhaustive. Chan and Harris state that making individuals more prosocial cannot count as making them morally better persons since true moral agents do not only perform actions with morally right consequences; they are moreover (reflectively) concerned about what the morally right action is (Chan & Harris 2011). Other philosophers also hold this view (Korsgaard 2006). To truly enhance an individual morally simply requires making that individual more morally reflective.

Even though Chan and Harris do not explicitly state it, we take their concern to relate to another demand requiring that if making an individual more prosocial is to count as making her a morally better person, then the relevant individual must have the right motivation(s) to be prosocial. This relates to empirical work on prosocial behaviour. In research on altruism, it is often stressed that individuals can have both selfish and unselfish motivations for their prosocial attitudes and behaviour (Batson 1991). An individual might be altruistic because it brings her social reward (a selfish motivation concerning for example social reputation and coalition building) or out of genuine unselfish concern of the other individual (a genuine motivation to benefit the well-being of others). One might claim that only the unselfish type seems fit as qualifying for making a person a morally better person.

Prosociality as moral impairment. Another category of warning stresses that being more prosocial might actually hinder you in being a morally good person and prevent you

from doing the morally right thing. Sparrow points to the case that making a judge more emphatic (empathy is taken to be a prosocial state or trait) might hinder the judge in doing the morally right thing (Sparrow 2014). Despite clear and overwhelming evidence, the judge might declare a person non-guilty due to strong emphatic relating and, hence, refrain from doing the morally right thing (judging in accordance with the law and the available evidence).

In addition, Chan and Harris (2011) point to a study by Crockett et al. (2010) in which a group of individuals were administered citalopram (a selective serotonin reuptake inhibitor). Researchers think that citalopram makes individuals more prosocial in certain respects. This prosocially enhanced group showed two tendencies. First, they were less willing to kill one person to save five other persons in a personal moral dilemma task. Second, they were less likely to reject unfair offers in a game called the Ultimatum Game.[9] However, such tendencies may not always be morally desirable. As Chan and Harris (2011) point to, in some real-life situations the morally right thing is indeed to inflict harm (e.g., we may hinder the death of a hundred innocent airplane passengers by harming a terrorist) or to reject and rebel against unfair and unjust transactions (e.g., we should not accept unfair payment and exploitation of workers).

In addition, prosocial effects of other bio-medical drugs might be taken to show morally undesirable effects. For example, studies show that the compound oxytocin facilitates certain tribal or ethnocentric prosocial effects: i.e., the compound appears to make individuals more prosocial to their in-group members, but more anti-social to out-group members (De Dreu et al. 2011; Shalvi & De Dreu 2014. And although these results are controversial in some respects see also Bartz et al 2011; Lane et al. 2015, 2016). Such 'tribal' prosocial effects seem morally problematic in many respects; after all, they occur to affirm ethnic and cultural pre-defined groups and might be viewed as generating racism and other morally undesirable dynamics.

Prosociality as unintelligent. Related to the above considerations, other authors have stressed that making people more prosocial might cause them to become less intelligent. Shook has argued for such: he states that making people more "emphatic, altruistic, or trusting" (all prosocial traits) may make them "dangerous fools, or worse" (2012, 11). This seems relevant. After all, some contexts are simply not fit for prosociality in certain

9. The Ultimatum Game has the following structure. One individual (or a computer) makes an offer to another individual on how to share some sum between them. The individual receiving the offer can either choose to accept the offer (which means that the sum is shared between them according to the offer) or reject it (which means that neither individuals gets any part of the sum).

respects since such attitudes and behaviour would be foolish or naïve (for example, think of an individual being prosocial towards other manipulative and exploitive individuals).

In relation to this, another concern may be that even if contexts are fit for pro-social relating, then they often demand different ways of realising one's prosociality. Think of the difference between being prosocial toward a child who is lost in a supermarket and acting prosocially towards an adult person who has lost a relative; such situations call for different ways of intending to benefit the other individual. Further, consider how situations might develop such that they shift in respect to what way of prosocial relating that is fitting. During only a single evening, at one point in time it may be appropriate to be prosocial to your friend by constructively helping her with her paper and at another time by having caring attitudes toward her. These cases circle around the intuition that if being more prosocial is to be morally desirable, such prosociality has to be flexible and sensitive to contextual factors. This increased prosociality must be intelligent in other words.

2.4 Six Necessary Conditions

For the sake of conceptual clarity, we propose to order the above categories of objections further into a set of six *necessary* conditions stating the boundaries for when an increase in prosociality can count as a moral enhancement. This paper does not argue that this set also constitutes sufficient conditions for morally enhancing an individual through prosociality. This corresponds to the ambition of the main claim of the paper. This claim is not that psilocybin facilitated prosocial effects actually constitutes moral bio-enhancement. Instead, the claim is that psilocybin is worthy future research attention because the prosocial effects, it facilitates, satisfy the common worries in the philosophical literature (as we have expressed them in the following necessary conditions). Again, it is important to stress that this paper does not outline these conditions to argue that they are correct. We outline these conditions to represent popular views in the philosophical literature. More formally we propose to present the challenge against performing moral bio-enhancement through prosocial effects as follows.

> Changing an agent, A, to become more prosocial toward some individual or group, G, (i.e., A has increased motivation to benefit G and is therefore more prone to behave accordingly) counts as morally enhancing A, only if

(i) *Social role*. A does not occupy a social role in which increased paradigmatic prosocial attitudes or behaviour is generally seen as disruptive for properly carrying out that role.

(ii) *Motivation*. A's increased prosociality is, at least partially, based upon a genuine unselfish concern for G—or because A take it, by principle, to be morally correct to be prosocial to G.

(iii) *Reflection*. A has, at some point, exercised sufficient moral reflection in relation to this increase in prosociality towards G; meaning that this increase in prosocial attitude and behaviour has some relevant causal relationship to A's reflection upon what she morally ought to do.

(iv) *Intelligence*. A's increased prosociality to G is sufficiently intelligent.

(v) *Moral traits*. A's increased prosociality to G does not decrease another morally desirable trait of A to an unacceptable degree.

(vi) *Tribalism*. A's increased prosociality to G must not be connected with an improper increased antisocial attitude to members outside of G.

Let us make some general comments on this outline. First, some of these requirements might be overlapping—one might reduce one of them to another (as the following elaboration might show, candidates could be to reduce (ii) to (iii), or (i) to (v)). However, no reduction will be attempted here and we shall discuss each condition separately.

Second, depending on one's general moral outlook, one might think that some of the requirements are irrelevant since they do not rest on any convincing moral consideration(s). Take for example a utilitarian minded person: she would probably find (ii) and (iii) to be only of instrumental importance. Likewise, one might imagine that a Kantian would stress that the only factor determining the moral character of a person is whether she acts out of duty or respect to the moral law (this would make (ii), or the second disjunct in it, the only legitimate condition). This taps into the profound disagreement mentioned in Section 2.3 on what characterises the morally good person. As already mentioned, this paper does *not* take position on which of these conditions that are in fact plausible by further examination (relative to different moral theories and systems). We include them all here to represent the common philosophical worries about

performing moral bio-enhancement through prosociality as comprehensively as possible. Neither do we wish to exclude or favour any particular major moral system by the above set of conditions; we wish to include them all in the set of requirements.

Let us now elaborate on the six requirements individually. Concerning (i), this requirement states that for A to be a morally better person by having increased prosociality to some individual or group, A must not occupy a social role in which paradigmatic prosociality is generally seen as problematic. Following Sparrow (2014), being a judge might be an example of such a role. Depending on the reader's view of the proper function of various roles, she might take specific leader positions, political roles, or military functions likewise to be such social roles in which increased prosociality is generally unfitting.

Concerning (ii), this condition demands that A's increased prosociality to an individual or group must (at least partially) include an unselfish other-concerned motivation or a principle-driven motivation to act prosocially out of the intention to follow some moral rule or principle demanding this. Again, some utilitarian minded people might find this requirement irrelevant. Others, such as psychological egoists, might find it unrealistically demanding (since all behaviour is ultimately selfish according to this view). Yet, based on the review of the previous part, the condition is included here.

In relation to (iii), this requirement demands that A's increased prosociality is manifested in connection to some sufficient level of moral reflection exercised by A herself. This requirement is included to deal with the view, described above, that 'full-bodied' moral agents reflect upon what they morally ought to do (Chan & Harris 2011). This view of what constitutes a moral agent is not easy to translate into a clear formulated condition. Nonetheless, we take one plausible implication of such a view to be the claim that for any prosocial tendency in an agent to count as making that agent a morally better person, this tendency must (at some point in time) have been reflected upon and, in relation, 'reflectively endorsed' by the agent herself. Further, such moral reflection must stand in a causal relationship to the increase in prosociality such that this increase is at least partially caused by the reflection at some point in time. The relevant reflection must simply not be fully post hoc to the prosocial attitude or behaviour. This description should be elaborated upon in the following way.

First, concerning the causal aspect of how the prosocial tendency of the individual arises, we take it to be plausible to read (iii) as stating that the prosocial tendency must either initially arise at least *partially* due to moral reflection; or that if moral reflection plays no causal role in the initial arising of the tendency in the individual, the tendency must be maintained by moral reflection of the individual.

Second, concerning another causal and temporal aspect, we take it to be *too* strong to read (iii) as demanding that: if an act, C, made by an agent, A, in a situation, S, is to count as making A a morally better person, C must be proximately caused by explicit moral reflection exercised by A in S. Take an example: B falls of a ship into the sea under a terrible storm, by 'pure instinct' or 'automatic processes' A jumps into the sea and saves B without any preceding proximate moral reflection. Condition (iii) does not have to deny this action as counting positively in the evaluation of the moral character of A. Instead, it may be more plausible to read (iii) as only demanding that A, at some point in time, must have reflected upon what she morally ought to do and, upon reflection, endorsed the view that one morally ought to risk one's life to save the lives of other people in some situations (or a similar view). This distant reflection should then play some more ultimate causal role and influence an agent's more instinctive behaviour.

Third, concerning the style of moral reflection, one might wonder what (iii) exactly demands. To demand reflection similar to that of academic moral philosophy seems too much; very few humans engage in such a style of reflection. Hence, let (iii) only demand the type of common moral reflection that most people undergo at some point in their lives. We take such reflection to circle, more or less explicitly, around questions such as 'what ought I morally to do?' and 'what is a morally good person?' but without any reliance on philosophical theories.

Requirement (iv) is not easy to flesh out, but the requirement is an incorporation of many of the worries around prosociality as being foolish and unintelligent. As it is simply stated, (iv) demands that the increased prosociality must be intelligent (Earp 2018). By intelligence we generally mean the ability to engage flexibly and context-sensitive with one's surroundings to achieve a goal by executing heterogeneous behaviour and adaptive modifications (this is a common textbook description of intelligence) (for example in Stanford et al. 2017, 410; Wyatt 2017, 57). Following this, the increased prosociality must be regulated through a capacity to engage flexibly with the present circumstances.

Yet, it is still unclear what level of intelligence "sufficiently" refers to in (iv). Since morality is often, if not always, primarily a social matter (in the basic sense that it involves living with others) let us focus on social situations. Thus, let us take 'sufficiently' to refer to the level that a person of average social intelligence operates on in relation to social situations. Such a person can, by a certain accuracy, understand different norms in different communities, register changes in social dynamics, map hierarchies in various settings, observe coalition-building, generate predictions about both the short- and long-term social consequences of an action, and form beliefs with a certain accuracy about the motives, emotions, and personalities of different individuals. Following this, let us say

that (iv) demands the increased prosociality to be processed and regulated through such socially intelligent capacities if it is to count as moral enhancement.

In addition, requirement (v) demands that the increased prosociality must be constituted "without reducing another morally desirable trait of A to any unacceptable degree". This requirement is motivated by the warnings on prosociality as moral impairment. What traits we include as morally desirable, how we measure them, and what "to an unacceptable degree" precisely refers to are questions beyond the scope of this paper. However, it is important to note that (v) does allow for decreasing a morally desirable trait to some degree; something that we might be inclined to do if this decrease is associated with an increase in another morally desirable trait (i.e., we might accept different trade-offs). We will return to (v) in Section 3.2.

Concerning requirement (vi), this condition demands that if an individual's increased prosociality is to count as making her a morally better person, this prosociality must not be tribal of nature meaning that it involves an increased motivation or readiness to disadvantage out-group members. Note, however, that requirement (vi) does allow that an individual's increased prosociality is limited only to her own in-group (i.e., that the relevant individual or group is made up only of in-group members), as long as this does not involve any increased antisocial attitudes to out-group members.

With these conceptual issues clarified, the following Section 3 suggests that the prosocial effects facilitated by psilocybin likely satisfy this set of necessary conditions.

3. Psilocybin and Prosociality

This section discusses the prosocial effects of psilocybin and their relation to the six necessary conditions stated in Section 2.4. When we talk about the 'prosocial effects of psilocybin' we refer to the case that psilocybin increases the prosocial attitudes, and hereby proneness to prosocial behaviour, of individuals. Importantly, this section considers the *lasting* prosocial effects of psilocybin: i.e., the prosocial effects that pertain after the compound of psilocybin is no longer active in the individual. This is to be contrasted with *acute* prosocial effects of psilocybin, which are effects that only occur while psilocybin is directly neurobiologically active in the individual (such acute effects are taken to occur within twenty hours of consumption of psilocybin) (Mason et al. 2019).

Initially, we must stress that it is widely recognized in research that the following three main variables determine the effects of consuming psilocybin and other psychedelic substances (Fadiman 2011; Zinberg 1984; Studerus et al. 2012).[10]

Dosage of specific substance. In general, the psychological effects of psychedelics are dose-dependent (Milliére et al. 2018; Dolder et al. 2016). If we were to carry out moral bio-enhancement by psilocybin we would have to hold precise knowledge of dose-dependent lasting effects of the substance in relation moral cognition.

Set. Another variable is the mind-set, or simply set, of the individual using the substance (Fadiman 2011, 16). The term 'set' encapsulates both the current psychological state (e.g., is the person oriented towards developing her moral character?) and the more standing traits (e.g., is the person generally emphatic?) of the individual consuming the substance. Individuals with different (mind)sets will undergo different changes by consumption of the same dose of psilocybin.

Setting. The term setting refers to the surroundings in which the substance is consumed—these surroundings being both the physical (e.g., is the location warm and inviting?) and the social environment (e.g., is the atmosphere trusting?) (Richards 2015; Carhart-Harris et al. 2018a; Hartogsohn 2016).

Dosage, set, and setting are important for the discussion of this paper for several reasons. First, notice that *no* study has been conducted in which the set and setting was purposely primed or designed to facilitate moral enhancement. With this in mind, we might be optimistic about how a purposefully 'moral enhancement oriented' designed dosage, set, and setting administering could facilitate even bigger and more lasting prosocial effects than current evidence suggests. Second, as researchers often stress, psilocybin 'only' facilitates certain psychological effects, such as prosociality, in an interplay with the set and setting (Richards 2015). The following parts refer to these three variables together as DSS-conditions.

3.1 Lasting Prosocial Effects of Psilocybin

Although there are several interesting studies on the potent acute prosocial effects of psilocybin (Porkorny et al. 2017; Gabay et al. 2018), the following is dedicated to studies touching upon the lasting effects.

10. Psychedelics are typically characterised as serotonin 2A receptor (5-HT_{2A}R) agonists with distinct psychological effects) (Johnson et al. 2019; Carhart-Harris & Nutt 2017; Roseman et al. 2018, 974).

A seminal study by Griffiths et al. (2006) administered 30mg/70kg psilocybin to thirty individuals in a double-blind study. Among many interesting things, the study found that two months after the consumption of psilocybin, participants reported themselves to have had significant positive changes in their mood and behaviour (ascribing it to the psilocybin-experience). Importantly, such effects included a general positive behaviour change, more positive attitudes to oneself and others, and general positive social and altruistic changes. Moreover, community members (e.g., friends or family to the participants) who had regular contact with the participants also reported that the participants had undergone a positive change in behaviour and mood, including a positive social change. Griffiths et al. (2008) further investigated whether these positive effects were present fourteen months after the study and found that most participants (61%) reported that the positive change in mood and behaviour had lasted to a moderate or even extreme degree. Among this lasting positive change, positive social and altruistic changes were present. Further, many of the participants reported that the psilocybin-trip had facilitated insights about love and empathy (these insights are not specified in-depth in the study).

Evidence strongly suggests that the prosocial related effects of psilocybin are dose-dependent. In another study by Griffiths et al. (2011) participants were either given a placebo drug or 5, 10, 20, or 30 mg psilocybin per 70 kg body weight. All participants administered psilocybin reported to have had positive changes in behaviour including positive changes in attitudes to life, altruistic/positive social change, increased empathy, and better connection to other individuals—yet, the effects were most significant by 30mg/70kg administering. These effects were lasting: after fourteen months subjects still reported them to be present. Moreover, once again, reports from community members close to the participants also stressed the positive change in behaviour and mood.

More recent studies cohere with the above results. Agin-Libes et al. (2020) found that psilocybin assisted psychotherapy had positive effects on social behaviour and relationships even 4,5 years after the use (in general, 71-100% of participants reported that the psilocybin assisted therapy had had overall positive life changes lasting this long). In a web-based survey study of 886 subjects who had participated in a psychedelic group session, Kettner et al. (2021) concluded that the psychedelic ingestion in a social setting led to enduring pro-social effects such as increased interpersonal tolerance and social connectedness in the majority of the study-participants.

In a much more limited time scope, Mason et al. (2019) found that psilocybin increased emotional empathy a week after consumption. This picture is affirmed by multiple other recent studies: they affirm that many of the attractive psychological

effects of psilocybin, including prosocial effects, are lasting in multiple months (see for example Griffiths et al. 2018), and even up to multiple years, although individuals might need more than one consumption of psilocybin to gain such effects (Studerus et al. 2011; Erritzoe et al. 2018; MacLean et al 2011; Lerner & Lyvers 2006; van Mulukom et al. 2020).

At last, worth noticing, large population studies show that use of psychedelic substances, for example psilocybin, is associated with decreases in antisocial behaviour such as assaults and partner violence (also when covariates are controlled for) (Hendricks et al. 2018; Thiessen et al. 2018). These studies suggest that psilocybin facilitates such lasting prosocial effects (and decreases antisocial behaviour) by generally increasing the capacity for emotion regulation among individuals (Thiessen et al. 2018; Young 2013, 78). This is interesting since emotion regulation is generally seen as necessary for competent moral cognition and appropriate behaviour (Zhang et al. 2017).

With the above overview in mind, there is a strong case to make that psilocybin, under the appropriate DSS-conditions, facilitates lasting prosocial effects in individuals. Several other authors than us also suggest this (Ahlskog 2017; Earp 2018; Tennison 2012; Haidt 2012, 265; Pollan 2019, 273). It is also very important to stress that the increased prosociality of subjects who consume psilocybin usually has a very broad segment: subjects not only have increased prosocial attitudes to their close relatives, but to human beings in general and even extending to non-human organisms (see for example Griffiths et al. 2006, 2008, 2011, 2018; Studerus et al. 2011; Carhart-Harris et al. 2018b, and for increased concern for non-human organisms and nature as a whole, see Forstmann & Sagioglou 2017; Kettner et al. 2019).

3.2 Examination by the Six Necessary Conditions

While we have no direct evidence to determine whether the prosocial effects facilitated by psilocybin satisfy the six necessary conditions stated above, we can indeed make the following examination.

First, considering (i): it requires that the relevant subject of the moral bio-enhancement does not occupy a specific social role in which an increase in paradigmatic prosociality is generally seen as problematic. Simply selecting a proper subject as target for the enhancement would satisfy this condition.

Recall requirement (ii): it demands that the change to more prosocial relating to other people must have an element of genuine unselfish other-concern. Studies show that around two/thirds of the participants under high-dose psilocybin consumption rank their

experience as either among the five most meaningful or the most meaningful experience of their lives (Griffiths et al. 2006, 276). Likewise, more than 60% of participants in a related study rated the experience to be among the top five spiritual experiences of their lives (Griffiths 2008, 627). Moreover, in another study by Griffiths et al. (2011, 661) participants had likewise profound feelings of unity, sacredness, and connectedness to the world. At last, as already mentioned, many participants under high-dose psilocybin have so-called 'mystical experiences' in which they have a sense of losing their self or ego, and their ego-centric orientation is decreased (Lebedev et al. 2015).[11] In such states, individuals report to feel more connected to themselves, other people, and the world.[12]

With these strong and life-changing experiences of unity, connection, and loss of ego-centric orientation we take it to be reasonable to claim that the prosocial effects of psilocybin do at least partially have a genuine unselfish other-oriented motive of care and concern for other people (a view Gabay et al. (2018, 8236) also support). Participants generally seem to be less egocentrically motivated and more turned towards both their human and non-human surroundings. Moreover, qualitative studies confirm that a greater degree of other-concern is a common result of psilocybin consumption (Belser et al. 2017).

Considering (iii), it demands that "sufficient moral reflection must be connected to the change" in prosociality. Studies do seem to affirm that, under the appropriate DSS-conditions, the prosocial effects of psilocybin can be said to have some moral reflective dimension. Recall that many participants under high-dose of psilocybin—especially those reaching the 'mystical experience'—reported to gain deep insights into their own psychology, the workings of the world, and often philosophical and religious questions such as how one should relate to one's surroundings (Griffiths et al. 2006, 2008, 2011). We think such changes in perspective and the associated increased understanding connect to moral questions on how one ought to live. Further, these changes are generated by highly rich mental activities that draw on memory, emotions, ideals, imperatives, and a broad range of components found in deep thoughtful practice. This idea is supported

11. Importantly, the expressions 'self' and 'ego' are not trivial to specify and they are subject to great controversy both in philosophy and cognitive science. Here, we refer roughly to the conceptualisations of these expressions as done in the Ego-Dissolution Inventory (see: Nour et al. 2016, 269). See also Milliére et al. (2018) for an illuminating discussion of these matters.

12. 'Connectedness' has become a key term in psychedelic science and some authors propose that the general therapeutic effects of psychedelics are always established and mediated by connection of individuals to the relevant surroundings. See Watts et al.(2017) and Carhart-Harris et al. (2018b).

by multiple qualitative studies; they point to the case that psilocybin experiences often involve morally reflective dimensions or examination of one's own values (including values of a moral or ethical character) (see Belser et al. 2017; Watts & Luoma 2020; Watts et al. 2017; Swift et al. 2017; Noorani et al. 2018; Hartogsohn 2018, 129).

Next, consider requirement (iv): it demands that the individual must relate more prosocially to other people in a "sufficiently intelligent way". We think that the two below considerations suggest that psilocybin facilitated prosocial effects do satisfy this requirement.

First, recall the three studies by Griffith et al. (2006, 2008, 2011) from Section 3.1. Participants in all these studies reported that psilocybin, especially under high-dose administering, had made a *general* positive contribution to their behaviour and mental life (an assessment in which community observers agreed). These results seem to suggest a positive relation between psilocybin and general intelligence in relating to themselves and others—this increased intelligence being a lasting effect. At least, the studies do in no way indicate that participants became less competent in general intelligent behaviour.

Second, studies show that psilocybin does increase, as a lasting effect, the general psychological flexibility of individuals (i.e., the ability to navigate adaptively in the present moment) and facilitates that individuals relate cognitively and emotionally less rigidly to their environment (Davis et al. 2020).

With these studies in mind, we think it is reasonable to be optimistic about satisfying (iv): under the right DSS-conditions, psilocybin seems to contribute positively to a person's general cognition and psychological flexibility. This makes sense since psilocybin is thought to work in a holistic way, which means that the effects of this substance are generated by and integrated in the larger general intelligence of the person (Carhart-Harris et al. 2014).

Though, one consideration must be mentioned here. Multiple studies show that psilocybin acutely biases emotion recognition such that it disrupts the ability to recognise negative emotions in other individuals (lysergic acid diethylamide, LSD, has similar effects) (for a review of these effects see Rocha et al. 2019). These results might count against the claim that (iv) is satisfied, since one might take the ability to recognise negative emotions to be crucial for moral competent cognition and behaviour. Though, notice that the disruption of recognition of negative emotions has primarily been shown to be an *acute* effect of psilocybin—whether it is a lasting effect seems to be less clear. Moreover, even if these effects appeared to be lasting, they had to be sufficiently big if they were to count as direct disruptions of the intelligence of the individual.

Further, recall requirement (v): it demands that the increase in prosociality must not reduce "another morally desirable trait of A to any unacceptable degree". To our knowledge, no study can provide us with a satisfying answer to this question. The closest we get are two recent studies. Porkony et al. (2017) investigated whether psilocybin had any *acute* effects on moral decision-making—they found that psilocybin had no significant effects on moral decision-making as measured by moral dilemma tasks. Perhaps more relevant, a study by Gabay et al. (2018) investigated the effects of psilocybin on behaviour in the Ultimatum Game. Except from one participant, all individuals in the psilocybin condition accepted all fair (50% share offered) and hyper-fair offers (80-90% share offered). Concerning unfair offers (20-30% share offered), psilocybin reduced the rejection rate of such offers (i.e., participants in the psilocybin-condition accepted more unfair offers).

Of the results obtained by these studies, only the reduced rejection rate of unfair offers might be understood as evidence for a decrease in a morally desirable trait. As Gabay et al. (2018) discuss, punishing unfair offers by rejection might be understood as altruistic punishment (altruistic punishment being the costly punishment of letting go of own reward to punish a norm violation, here the violation of the norm of fair sharing). Altruistic punishment might be understood as connected to some sense of justice—a sense we would count as a morally desirable trait. With this in mind, one might claim that (v) is not satisfied.

However, even if we accept this understanding of the results, one could reply in the following two ways. First, the mentioned study investigated the *acute* effects of psilocybin on moral decision-making, while the present discussion considers whether the lasting prosocial effects of psilocybin are relevant for moral enhancement. Decreases in altruistic punishment might only be an acute effect of psilocybin.

Second, as an explanation of the reduced rejection in the psilocybin condition, Gabay et al. (2018) propose that this behaviour is caused by an increase in other-concern in individuals in this condition (the explanation being that subjects in the psilocybin condition would not reject unfair offers since this would leave the other participant empty-handed). This increased other-concern might count as a morally desirable trait too, which would imply that we would have to weigh the decrease in 'altruistic punishment dispositions' with the increase in 'other-concern' before making the call that (v) is not satisfied. In such a weighting, we might believe that in relation to many real-world social scenarios a trade-off of having increased other-concern and decreased tendency to altruistic punishment is morally attractive. If we restrict ourselves to only performing moral bio-enhancement through psilocybin on individuals who do *not* hold roles or

positions in institutions in which altruistic punishment is crucial (a matter that really concern condition (i)), we could further justify this reply and weighing.

Though, the above issues shall not be discussed further here. Determining whether (v) is satisfied would demand discussions of what traits were to count as morally desirable, how we are to measure such traits, and what trade-offs to accept. However, since we have seen how psilocybin makes very positive general changes to a person's psychology, we might be optimistic about its contribution to a variety of morally desirable traits too. Also, recall that no study has worked with purposefully designed and primed DSS-conditions for moral bio-enhancement—this might as well fuel optimism about satisfying (v).

Then consider condition (vi): it demands that the relevant person's increased prosocial attitude must not be connected with an increased antisocial attitude to out-group members. Based on the previous mentioned studies, it does seem reasonable to think that the prosocial effects of psilocybin are not of tribal nature. First, as also mentioned in Section 3.1, the target of the increased prosocial attitudes appears very encompassing; meaning that participants have increased prosocial attitudes to human beings in general, and also often even to non-human animals and nature in general (Watts et al. 2017; Carhart-Harris et al. 2018b). Second, no study has reported that psilocybin is associated with increased anti-social tendencies to another individual or group. This absence of evidence is of course not direct evidence for psilocybin not being associated with anti-sociality, but it stresses that psilocybin so far does not have a research record involving anti-sociality.

Yet, some qualitative research indicates that psilocybin *can* generate tribal prosocial effects. As Langlitz (2020) notices, some contemporary far right-wing groups use psychedelics, such as psilocybin, to strengthen in-group prosociality and out-group anti-sociality. Further, Langlitz also points to anthropological studies showing that different Amazonian communities have used psychedelic compounds to increase antisocial and militant attitudes to out-group members (Dobkin de Rios 1984). These are important considerations. However, no contemporary scientific psilocybin study (conducted by researchers and professional clinicians or councillors) has reported that psilocybin facilitates such tribal forms of prosociality. Therefore, it seems reasonable to suggest that tribal forms of psilocybin facilitated prosocial effects, as well as non-tribal forms, are determined by the relevant DSS-conditions (Langlitz (2020) draws the same conclusion). With this in mind, we conclude that requirement (vi) can be satisfied when psilocybin consumption happens under the appropriate DSS-conditions.

With the above examination in mind, we find it reasonable to tentatively and moderately optimistic about the case that psilocybin facilitated prosocial effects, under the appropriate DSS-conditions, will satisfy the six necessary conditions.

3.3 The target of Prosociality

The reader might think that one crucial issue has been neglected in the above sections: namely, the more specific identity of the individual or group being the target of the increased prosociality. The reader might think that the identity of the individual or group is crucially decisive for whether increases in prosociality are morally attractive (e.g., an agent exhibiting prosocial attitudes and behaviour towards groups with suspect or wrong intentions, such as the execution of unjust violence, cannot count positively to the moral goodness of that agent). This issue of identity is obviously relevant and deserves a brief elaboration here.

One way to respond to this worry would be by stressing that the above issue of identity is in fact incorporated in multiple of the above six necessary conditions. For example, condition (iii) demands that moral reflection must play some causal role in the relevant increased prosociality. One could further elaborate upon (iii) by stressing that such moral reflection *must* involve reflections upon what constituted proper targets of prosociality. To satisfy (iii), the agent simply had to substantially reflect upon what individuals and groups she should be motivated to benefit.

Another conditions potentially dealing with the issue of identity would be that of (iv). As the outline and elaboration of condition (iv) stresses, the increase in prosociality must be intelligent in the way that it must be integrated in the agent's general ability of social cognition. One could take this general ability to involve the capacity to competently distinguish between proper and improper targets of prosociality. To satisfy (iv) the agent's increased prosociality had to be integrated into her general ability to distinguish proper from improper targets of benefit.

Last, a third condition dealing with the issue of proper targets of prosociality would be (v). Condition (v) demands that no other moral traits must suffer to an unacceptable degree due to the increase in prosociality. One could argue that this would involve the ability to distinguish proper from improper targets of benefit. One way to further characterise this ability would be to label it as a form of sense of justice (the sense of who deserves benefit). Condition (v) would hereby be taken to implicitly demand that the target of prosociality was not morally improper.

Yet, this implicit treatment of the issue of identity might not satisfy the reader and she might respond that the issue is way too important to be treated in such an implicit manner, by implementing it in the formulations of other conditions. The reader might alternatively stress that adding a seventh conditions to the overall set of necessary conditions is demanded to more explicitly treat the issue of identity. This condition could be formulated as below.

(vii) *Target*. G must be a proper target of prosociality.

The formulation of this condition is obviously under-specified as it stands (e.g., in relation to the morally laden nature of the qualification 'proper'). This paper shall not offer a satisfying specification here. However, as we briefly propose below, there is good reason to overall believe that the targets of psilocybin facilitated prosocial effects can be controlled by DSS-conditions. This suggests that controlling for DSS-conditions enable control of the targets of increased prosociality facilitated by psilocybin. That is, recall from the above paragraphs that studies indicate that psilocybin facilitates both very global prosocial effects (ranging from relatives, to people in general, and even to non-human animals) and tribal prosocial effects involving out-group hostility. As also stressed in Section 3.2, this variance in the targets of prosocial effects strongly suggests that these effects are determined by the operating DSS-conditions. In the light of this, it does seem reasonable to be optimistic about satisfying (vii) by establishing fitting DSS-conditions (e.g., DSS-conditions in which a competent moral councillor or coach was present to guide the change in the moral psychology of the agent). Further, as we have argued above, since the conditions of (iii), (iv), and (vi) all have some important relation to the target of the increased prosociality, satisfying something like condition (vii) would most probably occur in tandem with the satisfaction of many of the other necessary conditions and in general with the appropriate DSS-conditions. Condition (vii) hereby fits very naturally with the content of many of the other conditions and their reliance on DSS-conditions.

4. Further Worries and Research

The optimism about satisfying the set of necessary conditions should however be seen as tentative in various respects. Numerous central issues remain unanswered. Further research on psilocybin for moral bio-enhancement would need to target several questions—questions especially pertaining to the DSS-conditions for moral

bio-enhancement. We think that the following list of questions highlights some of the aspects that needs to be addressed in further discussion of psilocybin for moral bio-enhancement.

(a) How sure are we that psilocybin actually has *real* behavioural and psychological prosocial effects? That is, since the research cited in this article relies so heavily upon self report, could we not simply interpret the results as suggesting that individuals *subjectively feel* more prosocial, without necessarily actually being more prosocial? Future research should navigate this question and aim to unpack the relationship between the subjective feeling and the more objective attribute of being more prosocial. However, as we have referred to in the previous sections, some studies certainly do suggest that individuals do not only feel more prosocial but that they actually also act more prosocially (see for example Section 3.1 and the reports by community observers).

(b) If psilocybin reliably facilitates real prosocial effects, how long-term are these effects? Studies would need to identify the temporal scope of prosocial effects of psilocybin and clarify whether these effects are fading over time and to what degree.

(c) What is the proper dose-frequency: how often and how strong doses of psilocybin should individuals take to generate the relevant prosocial effects? Although many researchers consider psilocybin to be a physiologically safe drug, this question should also consider potential side effects of repeated use (both psychological and physiological).

(d) Should the individual engage in any particular moral practice, like moral deliberation, to influence her mind-set before psilocybin consumption? Since the effects of psilocybin are so sensitive to DSS-conditions, such preparation would most probably be suitable—though, it is not obvious what it should involve.

(e) Under the psilocybin consumption itself, we expect that a councillor (or moral coach or supervisor) should most probably be

present (as in many scientific experimentations on the therapeutic effects of psilocybin) (Griffiths et al. 2018). Should the councillor excite the individual to engage with moral questions? Further, should the councillor in some way be a representative of the individual's moral tradition or system (for example, a priest, rabbi, Lama, imam, or a secular philosopher depending upon the identity of the subject)?

(f) As stressed by many psychedelic therapists, the long-term psychological effects of psychedelic experiences are very sensitive to how the individual integrates the psychedelic experience into her everyday life, and whether the individual has any social or cultural context to make such integration within (Bourzat & Hunter 2019; Eisner 1997; Saunders et al. 2000; Richards 2015, 2017). Hence, one might ask whether moral bio-enhancement by psilocybin should involve follow-up integration initiatives (e.g., exercises of moral deliberation or interviews) to ensure lasting prosocial effects and other morally relevant psychological effects.

Many of these questions stress the importance of DSS-conditions, as this article has generally done. Some readers might wonder whether the dependence upon certain DSS-conditions makes the prosocial effects of psilocybin fragile in the sense that they are so context dependent that unattractive side-effects and other unexpected effects will inevitably occur upon broad use (making psilocybin too risky a bio-medical substance to use for moral bio-enhancement) (Fabiano 2020). This is obviously a reasonable worry and it applies to most, if not all, feasible proposals of a bio-medical or neuro-technological tool for moral bio-enhancement. We offer no general and satisfying answer to this here. We can only refer to the studies outlined in Section 3.1 and stress that under the DSS-conditions applied in scientific research, psilocybin reliably and consistently facilitates prosocial effects of a relatively common character—appearing to avoid unexpected side-effects. This provides hope that controlling for DSS-conditions will yield consistent changes in moral psychologies across individuals, populations, and trials.

A related yet distinct worry would be that seriously proposing psilocybin for moral bio-enhancement demands that one understands the general influence which psilocybin has on the workings of moral cognition (Dubljević & Racine 2017). If we do not understand these workings, we cannot come to know what the appropriate DSS-

conditions exactly are and hence control for them. This worry is also relevant. Though, moral cognition is a topic of great controversy: the literature contains multiple competing models of such cognition and none of them can claim the status of a standard model. Understanding the prosocial effects of psilocybin within these models would undoubtedly be of interest and would without doubt increase our understanding of how to control for the appropriate DSS-conditions. This matter is nevertheless beyond the scope of this paper. We have simply aimed to show that psilocybin, under appropriate DSS-conditions, facilitates prosocial effects that likely satisfy the outlined set of necessary conditions. How this fits into the discussions on the more specific workings of moral cognition and moral psychology is a matter for another occasion.

5. Conclusion

Section 2 of this paper provided a conceptual analysis of the meaning of 'prosociality' (more precisely, prosocial attitudes and behaviour). It further organised the common philosophical worries against performing moral bio-enhancement through prosocial effects into a set of six necessary conditions. Hereafter, Section 3 provided a review of the relevant research results on psilocybin. This section suggested that we have good reason to be tentatively and moderately optimistic about psilocybin, under the appropriate DSS-conditions, satisfying this set of necessary conditions. At last, Section 4 stressed the need for further research on DSS-conditions for moral bio-enhancement through psilocybin facilitated prosocial effects. In the end, this paper can be seen as an attempt to show that psilocybin for moral bio-enhancement is a matter worth of future research attention. Both philosophical and empirical work would be important in such future research. Philosophical work should explore the justification of prosociality as a target of moral bio-enhancement, and further scrutinize the requirements for increases in prosociality to count as moral enhancement. Empirical work should investigate both how psilocybin generally influences moral cognition and the proper DSS-conditions for its potential use in moral bio-enhancement.

References

Agin-Liebes, G. I., Malone, T., Yalch, M. M., Mennenga, S. E., Ponté, K. L., Guss, J., Bossis, A. P., Grigsby, J., Fischer, S., & Ross, S. 2020. "Long-term follow-up of psilocybin-

assisted psychotherapy for psychiatric and existential distress in patients with life-threatening cancer". *Journal of psychopharmacology* 34(2): 155–166

Ahlskog, R. 2017. "Moral Enhancement Should Target Self-Interest and Cognitive Capacity". *Neuroethics* 10: 363-73

Basurto, X., Blanco, E., Nenadovic, M., & Vollan, B. 2016. "Integrating simultaneous prosocial and antisocial behavior into theories of collective action". *Science Advances* 2(3): e1501220

Batson, C. D. 1991. *The altruism question: toward a social- psychological answer*. Hillsdale, NJ: Erlbaum

Batson, C. D., & Shaw, L. L. 1991. "Evidence for altruism: Toward a pluralism of prosocial motives". *Psychological Inquiry* 2: 107–122

Bartz, J.A., Zaki, J., Bolger, N., & Ochsner, K.N. 2011. "Social effects of oxytocin in humans: Context and person matter". *Trends in Cognitive Sciences* 15(7): 301–309

Beauchamp, T. L., & Childress, J. F. 2001. *Principles of biomedical ethics*. Oxford University Press, USA

Beck, B. 2015. "Conceptual and practical problems of moral enhancement". *Bioethics,* 29(4): 233-240

Belser, A. B., Agin-Liebes, G., Swift, T. C., Terrana, S., Devenot, N., Friedman, H. L., Guss, J., Bossis, A., & Ross, S. 2017. "Patient Experiences of Psilocybin-Assisted Psychotherapy: An Interpretative Phenomenological Analysis". *Journal of Humanistic Psychology* 57(4): 354–388

Boehm, C. 2012. *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York: Basic Books

Bourzat, F., & Hunter, K. 2019. *Consciousness Medicine: Indigenous Wisdom, Entheogens, and Expanded States of Consciousness for Healing and Growth*. North Atlantic Books;.

Caprara, G. V., Alessandri, G., & Eisenberg, N. 2012. "Prosociality: The contribution of traits, values, and self-efficacy beliefs". *Journal of personality and social psychology*, 102(6): 1289.

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R., & Nutt, D. 2014. "The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs". *Frontiers in human neuroscience* 8(20).

Carhart-Harris, R. L., & Nutt, D. J. 2017. "Serotonin and brain function: a tale of two Receptors". *Journal of psychopharmacology* 31(9): 1091–1120.

Carhart-Harris, R. L., Roseman, L., Haijen, E., Erritzoe, D., Watts, R., Branchi, I., & Kaelen, M. 2018a. "Psychedelics and the essential importance of context". *Journal of psychopharmacology* 32(7): 725–731.

Carhart-Harris, R. L., Erritzoe, D., Haijen, E., Kaelen, M., & Watts, R. 2018b. "Psychedelics and connectedness". *Psychopharmacology* 235(2): 547-550.

Chan, S. & Harris, J. 2011. "Moral enhancement and pro-social behaviour". *Journal of Medical Ethics* 37(3): 130-1..

Choi, J. K., & Bowles, S. 2007. "The coevolution of parochial altruism and war". *Science* 318(5850): 636-640.

Clark, M. S., & Taraban, C. 1991. "Reactions to and willingness to express emotion in communal and exchange relationships". *Journal of Experimental Social Psychology* 27(4): 324-336.

Clark, M. S., & Mills, J. 1993. "The difference between communal and exchange relationships: What it is and is not". *Personality and Social Psychology Bulletin* 19(6): 684-691.

Clark, M. S., & Mills, J. R. 2012. "A theory of communal (and exchange) relationships". In P. A. M. V. Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), Handbook of Theories of Social Psychology: Volume Two: 232-250. SAGE.

Crockett, M. J.; Clark, L., Hauser, M. D., Robbins, T. W. 2010. "Serotonin selectivity influences moral judgement and behaviour through effects on harm aversion". *Proceedings of the National Academy of Sciences of the United States of America* 107(40): pp. 17433-8.

Cummiskey, D. 1989. "Consequentialism, egoism, and the moral law". *Philosophical studies* 57(2): 111-134.

Curry, O. S., Mullins, D. A. & Whitehouse, H. 2019. "Is it good to cooperate? Testing the theory of Morality-as-Cooperation in 60 societies". *Current Anthropology* 60(1): 47-69.

Davis, A. K., Barrett, F. S., & Griffiths, R. R. 2020. "Psychological flexibility mediates the relations between acute psychedelic effects and subjective decreases in depression and anxiety". *Journal of Contextual Behavioral Science 15*: 39-45..

De Dreu, C. K., Greer, L. L., Handgraaf, M. J., Shalvi, S., Van Kleef, G. A., Baas, M., Ten Velden, F. S., Van Dijk, E. & Feith, S. W. 2010. "The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans". *Science*, 328(5984): 1408-1411.

De Dreu, C. K., Greer, L. L., Van Kleef, G. A., Shalvi, S., & Handgraaf, M. J. 2011. "Oxytocin promotes human ethnocentrism". *Proceedings of the National Academy of Sciences* 108(4): 1262-1266.

de Waal, F. B., & van Roosmalen, A. 1979. "Reconciliation and consolation among chimpanzees". *Behavioral Ecology and Sociobiology* 5(1): 55-66.

de Waal, F. 2006. "Morally Evolved: Primate Social Instincts, Human Morality, and the Rise and Fall of 'Veneer Theory'". In Macedo, S. & Ober, J. (eds.). *Primates and Philosophers.* Princeton: Princeton University Press: 1-75.

de Waal, F. 2008. "Putting the altruism back into altruism: the evolution of empathy". *Annual review of psychology* 59: 279–300.

Dobkin de Rios, M. 1984. *Hallucinogens: Cross-Cultural Perspectives. Prospect Heights*, Ill.: Waveland Press.

Dolder, P. C., Schmid, Y., Müller, F., Borgwardt, S., Liecht, M. E. 2016. "LSD Acutely Impairs Fear Recognition and Enhances Emotional Empathy and Sociality". *Neuropsychopharmacology* 41: 2638-46.

Douglas, T. 2008. "Moral Enhancement". *Journal of Applied Philosophy* 25(3): 228-45.

Dubljević, V., & Racine, E. 2017. "Moral enhancement meets normative and empirical reality: assessing the practical feasibility of moral enhancement neurotechnologies". *Bioethics*, *31*(5): 338-348..

Earp, B. D., Douglas, T., & Savulescu, J. 2017. "Moral Neuroenhancement". In L. Johnson & Rommelfanger, K. (Eds.). *The Routledge Handbook of Neuroethics*. Routledge.

Earp, B. D. 2018. "Psychedelic Moral Enhancement". *Royal Institute of Philosophy Supplement* 83: 415-39.

Earp, B. D., McLoughlin, K., Monrad, J., Clark, M. S., & Crockett, M. 2020. "How social relationships shape moral judgment". *PsyArXiv.*

Eisner, B. 1997. "Set, Setting, and Matrix". *Journal of Psychoactive Drugs* 29(2): 213–216.

Erritzoe, D., Roseman, L., Nour, M. M., MacLean, K., Kaelen, M., Nutt, D. J., & Carhart-Harris, R. L. 2018. "Effects of psilocybin therapy on personality structure". *Acta Psychiatrica Scandinavica*, 138(5): 368-378.

Fabiano, J. 2020. "The Fragility of Moral Traits to Technological Interventions". *Neuroethics*: 1-13.

Fadiman, J. 2011. *The Psychedelic Explorer's Guide*. Vermont: Park Street Press.

Forstmann, M. & Sagioglou, C. 2017. "Lifetime experiences with classical psychedelics predicts pro-environmental behaviour through an increase in nature relatedness". *Journal of Psychopharmacology* 31: 975–988

Gabay, A. S., Carhart-Harris, R. L., Mazibuko, N., Kempton, M. J., Morrison, P. D., Nutt, D. J., & Mehta, M. A. 2018. "Psilocybin and MDMA reduce costly punishment in the Ultimatum Game". *Scientific reports*, 8(1): 8236..

Gill, R. (Ed.). 2012. *The Cambridge companion to Christian ethics*. Cambridge University Press.

Griffiths, R. R., Richards, W. A., McCann, U., & Jesse, R. 2006. "Psilocybin can occasion mystical-type experiences having substantial and sustained personal meaning and spiritual significance". *Psychopharmacology*, 187(3): 268–83.

Griffiths, R., Richards, W., Johnson, M., McCann, U., & Jesse, R. 2008. "Mystical-type experiences occasioned by psilocybin mediate the attribution of personal meaning and spiritual significance 14 months later". *Journal of psychopharmacology* 22(6): 621–632.

Griffiths, R. R., Johnson, M. W., Richards, W. A., Richards, B. D., McCann, U., & Jesse, R. 2011. "Psilocybin occasioned mystical-type experiences: immediate and persisting dose-related effects". *Psychopharmacology* 218(4): 649–665.

Griffiths, R. R., Johnson, M. W., Richards, W. A., Richards, B. D., Jesse, R., MacLean, K. A., Barrett, F. S., Cosimano, M. P., & Klinedinst, M. A. 2018. "Psilocybin-occasioned mystical-type experience in combination with meditation and other spiritual practices produces enduring positive changes in psychological functioning and in trait measures of prosocial attitudes and behaviors". *Journal of psychopharmacology* 32(1): 49–69.

Haidt, J. 2012. *The righteous mind: Why good people are divided by politics and religion*. Penguin Vintage.

Hare, J. 2019. "Religion and Morality". In Zalta, E. N. (ed.). *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), URL = ‹https://plato.stanford.edu/archives/fall2019/entries/religion-morality/› .

Hartogsohn, I. 2016. "Set and setting, psychedelics and the placebo response: An extra-pharmacological perspective on psychopharmacology". *Journal of Psychopharmacology 30*: 1259–1267.

Hartogsohn, I. 2018. "The Meaning-Enhancing Properties of Psychedelics and Their Mediator Role in Psychedelic Therapy, Spirituality, and Creativity". *Frontiers in neuroscience*, 12: 129.

Helion, C., & Pizarro, D. A. 2015. "Beyond dual-processes: the interplay of reason and emotion in moral judgment". *Neuroethics* 11: 109-125.

Hendricks, P. S., Crawford, M. S., Cropsey, K. L., Copes, H., Sweat, N. W., Walsh, Z., & Pavela, G. 2018. "The relationships of classic psychedelic use with criminal behavior in the United States adult population". *Journal of psychopharmacology*, 32(1): 37-48.

Hilton, N. Z., Ham, E., & Green, M. M. 2018. "The roles of antisociality and neurodevelopmental problems in criminal violence and clinical outcomes among male forensic inpatients". *Criminal justice and behavior*, 45(3): 293-315.

Homiak, M. 2019. "Moral Character". In Zalta, E. N. (ed.). *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), URL = https://plato.stanford.edu/archives/sum2019/entries/moral-character/.

Jensen, K., Vais, A. & Schmidt, M. F. H. 2014. "The emergence of human prosociality: aligning with others through feelings, concerns, and norms". *Frontier in Psychology* 5(822).

Jensen, K. 2016. "Prosociality". *Current biology* 26(16): R748-R752.

Johnson, R. & Cureton, A. 2019. "Kant's Moral Philosophy". In Zalta, E. N. (ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), URL = ‹https://plato.stanford.edu/archives/spr2019/entries/kant-moral/›.

Johnson, M. W., Hendricks, P. S., Barrett, F. S., & Griffiths, R. R. 2019. "Classic psychedelics: An integrative review of epidemiology, therapeutics, mystical experience, and brain network function". *Pharmacology & therapeutics* 197: 83-102.

Joyce, R. 2007. *The Evolution of Morality*. Cambridge, MA: The MIT Press.

Kettner, H., Gandy, S., Haijen, E., & Carhart-Harris, R. L. 2019. "From Egoism to Ecoism: Psychedelics Increase Nature Relatedness in a State-Mediated and Context-

Dependent Manner". *International journal of environmental research and public health*, 16(24): 5147.

Kettner, H., Rosas, F., Timmermann, C., Kärtner, L., Charhart-Harris, R., & Roseman, L. 2021. "Psychedelic Communitas: Intersubjective Experience During Psychedelic Group Sessions Predicts Enduring Changes in Psychological Wellbeing and Social Connectedness". *Frontiers in Pharmacology* 12 https://doi.org/10.3389/fphar.2021.623985.

Kitcher, P. 2011. *The Ethical Project*. Cambridge, Massachusetts: Harvard University Press.

Korsgaard, C. 2006. "Morality and the distinctiveness of human action". *Primates and philosophers: How morality evolved*: 98-119.

Lane, A., Mikolajczak, M., Treinen, E., Samson, D., Corneille, O., de Timary, P., and Luminet, O. 2015. "Failed replication of oxytocin effects on trust: the envelope task case". *PLOS One* 10(9): e0137000.

Lane, A., Luminet, O., Nave, G., & Mikolajczak, M. 2016. "Is there a publication bias in behavioural

intranasal oxytocin research on humans? Opening the file drawer of one laboratory". *Journal of Neuroen-*

docrinology 28(4): 1–15.

Langlitz, N. 2012. *Neuropsychedelia. The Revival of Hallucinogen Research since the Decade of the Brain*. Berkeley: University of California Press.

Langlitz, N. 2020. "Rightist Psychedelia". Hot Spots, *Fieldsights*, July 21. https://culanth.org/fieldsights/rightist-psychedelia.

Lebedev, A. V., Lövdén, M., Rosenthal, G., Feilding, A., Nutt, D. J., & Carhart-Harris, R. L. 2015. "Finding the self by losing the self: Neural correlates of ego-dissolution under psilocybin". *Human brain mapping* 36(8): 3137–3153.

Lerner, M. & Lyvers, M. 2006. "Values and beliefs of psychedelic drug users: A crosscultural study". *Journal of Psychoactive Drugs* 38: 143–147.

MacLean, K. A., Johnson, M. W., & Griffiths, R. R. 2011. "Mystical experiences occasioned by the hallucinogen psilocybin lead to increases in the personality domain of openness". *Journal of psychopharmacology* 25(11): 1453–1461.

Mason, N. L., Mischler, E., Uthaug, M. V., & Kuypers, K. P. 2019. "Sub-acute effects of psilocybin on empathy, creative thinking, and subjective well-being". *Journal of psychoactive drugs*, 51(2): 123-134.

Miller, C. B. 2020. "Empirical Approaches to Moral Character". In Zalta, E. N. (ed.). *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), URL = https://plato.stanford.edu/archives/fall2020/entries/moral-character-empirical/.

Milliére, R., Carhart-Harris, R. L., Roseman, L., Trautwein, F., Berkovich-Ohana, A. 2018. "Psychedelics, Meditation, and Self-Consciousness". *Frontiers in Psychology* 9(1475).

Mosig, Y. D. 1989. "Wisdom and compassion: What the Buddha taught a psycho-poetical analysis". *Theoretical & Philosophical Psychology*, 9(2).

Neumann, C. S., Hare, R. D., & Pardini, D. A. 2015. "Antisociality and the construct of psychopathy: Data from across the globe". *Journal of personality* 83(6): 678-692.

Noorani, T., Garcia-Romeu, A., Swift, T. C., Griffiths, R. R., & Johnson, M. W. 2018. "Psychedelic therapy for smoking cessation: qualitative analysis of participant accounts". *Journal of Psychopharmacology* 32(7): 756-769.

Nour, M. M., Evans, L., Nutt, D., & Carhart-Harris, R. L. 2016. "Ego-dissolution and psychedelics: validation of the ego-dissolution inventory (EDI)". *Frontiers in human neuroscience (10).*

Parfit, D. 2011. *On what matters* (Vol. 1). Oxford University Press.

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. 2005. "Prosocial behavior: Multilevel perspectives". *Annu. Rev. Psychol.* 56: 365-392.

Persson, I., & Savulescu, J. 2012. *Unfit for the future: the need for moral enhancement*. OUP Oxford.

Pollan, M. 2019. *How to change your mind: What the new science of psychedelics teaches us about consciousness, dying, addiction, depression, and transcendence*. Penguin Books.

Pokorny, T., Preller, K. H., Kometer, M., Dziobek, I., & Vollenweider, F. X. 2017. "Effect of Psilocybin on Empathy and Moral Decision-Making". *The international journal of neuropsychopharmacology*, 20(9): 747–757.

Rai, T. S., & Fiske, A. P. 2012. "Beyond harm, intention, and dyads: Relationship regulation, virtuous violence, and metarelational morality". *Psychological inquiry*, 23(2): 189-193.

Raus, K., Focquaert, F., Schermer, M., Specker, J., Sterckx, S. 2014. "On Defining Moral Enhancement: A Clarificatory Taxonomy". *Neuroethics* 7: 263-73.

Regis Jr, E. 1980. "What is ethical egoism?". *Ethics*, 91(1): 50-62.

Richards, W. 2015. "11. Discipline and Integration". In *Sacred Knowledge*. New York: Columbia University Press: 119-125.

Richards, W. A. 2015. "Understanding the religious import of mystical states of consciousness facilitated by psilocybin." In J.H. Ellens and B. Roberts (Eds.). *The Psychedelic Policy Quagmire: Health, Law, Freedom, and Society*. Santa Barbara, CA, Denver, CO: Praeger: 139–144.

Richards, W. 2017. "Psychedelic Psychotherapy: Insights From 25 Years of Research". *Journal of Humanistic Psychology*, 57(4): 323–337..

Rocha, J. M., Osório, F. L., Crippa, J. A. S., Bouso, J. C., Rossi, G. N., Hallak, J. E., & Dos Santos, R. G. 2019. "Serotonergic hallucinogens and recognition of facial emotion expressions: a systematic review of the literature". *Therapeutic advances in psychopharmacology* 9: 2045125319845774..

Roseman, L., Nutt, D. J., & Carhart-Harris, R. L. 2018. "Quality of Acute Psychedelic Experience Predicts Therapeutic Efficacy of Psilocybin for Treatment-Resistant Depression". *Frontiers in pharmacology*, 8.

Saunders, N., Saunders, A., & Pauli, M. 2000. *In search of the ultimate high: spiritual experience from psychoactives*..

Schaefer, G. O. 2015. "Direct vs. Indirect Moral Enhancement". *Kennedy Institute of Ethics Journal* 25(3): 261-89.

Shalvi, S., & De Dreu, C. K. 2014. "Oxytocin promotes group-serving dishonesty". *Proceedings of the National Academy of Sciences,* 111(15): 5503-5507.

Shook, J. R. 2012. "Neuroethics and the Possible Types of Moral Enhancement". *AJOB Neuroscience* 3(4): 3-14.

Sparrow, R. 2014. "Egalitarianism and Moral Bioenhancement". *The American Journal of Bioethics* 14(3): 20-8..

Stanford, C., Allen, J. S. & Antón, S. C. 2017. *Biological Anthropology*, fourth edition. London: Pearson.

Strohminger, N., & Nichols, S. 2014. "The essential moral self". *Cognition*, 131(1): 159-171.

Studerus, E., Kometer, M., Hasler, F., & Vollenweider, F. X. 2011. "Acute, subacute and long-term subjective effects of psilocybin in healthy humans: a pooled analysis of experimental studies". *Journal of psychopharmacology*, 25(11): 1434-1452.

Studerus, E., Gamma, A., Kometer, M., & Vollenweider, F. X. 2012. "Prediction of psilocybin response in healthy volunteers". *PloS one* 7(2), e30800.

Swift, T. C., Belser, A. B., Agin-Liebes, G., Devenot, N., Terrana, S., Friedman, H. L., Guss, J., Bossis, A. P., & Ross, S. 2017. "Cancer at the Dinner Table: Experiences of Psilocybin-Assisted Psychotherapy for the Treatment of Cancer-Related Distress". *Journal of Humanistic Psychology* 57(5): 488–519.

Tennison, M. N. 2012. "Moral Transhumanism: The Next Step". *Journal of Medicine and Philosophy* 37: 405-416.

Thiessen, M. S., Walsh, Z., Bird, B. M., & Lafrance, A. 2018. "Psychedelic use and intimate partner violence: The role of emotion regulation". *Journal of psychopharmacology*, 32(7): 749-755.

Trautwein, F., Naranjo, J. R. & Schmidt, S. 2014. "Meditation Effects in the Social Domain: Self-Other Connectedness as a General Mechanism, in Schmidt, S. & Walach, H. (eds.). *Meditation—Neuroscientific Approaches and Philosophical Implications*. Cham: Springer: 175-99.

Twenge, J. M., Baumeister, R. F., DeWall, C. N., Ciarocco, N. J., & Bartels, J. M. 2007. "Social exclusion decreases prosocial behaviour". *Journal of personality and social psychology,* 92(1).

Van Kleef, G. A., Homan, A. C., Finkenauer, C., Blaker, N. M., & Heerdink, M. W. 2012. "Prosocial norm violations fuel power affordance". *Journal of Experimental Social Psychology* 48(4): 937-942.

van Mulukom, V., Patterson, R. E., & van Elk, M. 2020. "Broadening Your Mind to Include Others: The relationship between serotonergic psychedelic experiences and maladaptive narcissism". *Psychopharmacology*, 237(9): 2725–2737.

Walker, R. L., & Ivanhoe, P. J. (Eds.). 2007. *Working virtue: Virtue ethics and contemporary moral problems*. Oxford University Press.

Watts, R. Day, C., Krzanowski, J., Nutt, D., Carhart-Harris, R. 2017. "Patients' accounts of increased 'connection' and 'acceptance' after psilocybin for treatment-resistant depression". *Journal of Humanistic Psychology* 57(5): 520–564.

Watts, R., & Luoma, J. B. 2020. "The use of the psychological flexibility model to support psychedelic assisted therapy". *Journal of Contextual Behavioral Science* 15: 92-102.

Wikipedia 2021. Prosocial behaviour. Link: https://en.wikipedia.org/wiki/Prosocial_behavior. Accessed 04.08.21.

Wyatt, T. D. 2017. *Animal Behaviour—a very short introduction*. Oxford: Oxford University Press.

Young, S. N. 2013. "Single treatments that have lasting effects: some thoughts on the antidepressant effects of ketamine and botulinum toxin and the anxiolytic effect of psilocybin". *Journal of psychiatry & neuroscience: JPN* 38(2).

Zaki, J., & Mitchell, J. P. 2013. "Intuitive prosociality". *Current Directions in Psychological Science* 22(6): 466-470.

Zhang, L., Kong, M., & Li, Z. 2017. "Emotion regulation difficulties and moral judgment in different domains: The mediation of emotional valence and arousal". *Personality and Individual Differences* 109: 56-60.

Zinberg, N. E. 1984. *Drug, Set, and Setting: The Basis for Controlled Intoxicant Use*. New Haven: Yale University Press.

Journal of Cognition and Neuroethics

Does Physics Allow for Free Will?

Proposing a Novel Type of Psychophysical Experiments Testing the Multiverse Interpretation of Quantum Mechanics

Christian D. Schade
Humboldt-Universität zu Berlin

Biography

Christian D. Schade is a full professor at Humboldt University and holds the chair of Entrepreneurial and Behavioral Decision Making. His research mainly spans three fields: behavioral decision making and game theory, gender differences in decision making, and quantum decision making - including philosophical considerations on the existence of free will. He contributes to a better understanding of decision making in general (and what that actually is), of entrepreneurial as well as innovative decision making, as well as to a philosophical understanding of innovations. He is currently working on novel foundations and perspectives for the decision sciences. His research is mainly based on laboratory experiments, economic psychology and mathematical psychology, as well as quantum mechanics.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). January, 2022. Volume 8, Issue 1.

Citation

Schade, Christian D. 2022. "Does Physics Allow for Free Will? Proposing a Novel Type of Psychophysical Experiments Testing the Multiverse Interpretation of Quantum Mechanics." *Journal of Cognition and Neuroethics* 8 (1): 65–82.

Does Physics Allow for Free Will?

Proposing a Novel Type of Psychophysical Experiments Testing the Multiverse Interpretation of Quantum Mechanics

Christian D. Schade

Abstract

This contribution proposes a novel type of experiments that might be able to (a) test the multiverse interpretation of quantum mechanics against the standard interpretation, and (b) together with the theory of the clustered-minds multiverse (Schade 2018), might offer a proof of the existence of free will. The experiments are psychophysical in a novel sense, because, via top-down entanglement, consciousness is at the core of the measurement problem and influences the physical. At the core of the experiments are quantum-optical setups, together with a manipulation of the number, preferences, and state of information of the observers.

Keywords

Interpretation of Quantum Mechanics, Measurement Problem, Free Will, Consciousness, Top-down Entanglement, Clustered-minds Multiverse, Novel Psychophysical Experimentation, Quantum-optical Setups, Wave-particle Inequalities

Justifying Free Will via the Multiverse – And the Multiverse via Experimentation

Within the JCN special issue on free will (2015, vol. 3, issue 1), a plethora of diverging positions on free will have been presented, based on a conference held back in 2014 in Flint, MI. The papers are offering a great collection of contemporary treatments of the matter. Oftentimes, it becomes quite clear how essential the physical basis chosen or discussed by the respective author(s) is for the line of arguments presented by them.[1] Specifically, the physical basis chosen is often associated with the *opportunity* that is offered for free will. E.g., Cogley (2015) is defending current libertarianism against the potential problems arising from the assumed *indeterminism* – resulting from the standard interpretation of quantum mechanics – and Vihvelin (2015) makes the strong argument

1. This may partially be routed in the fact that the understanding of a concept that is more general than free will, i.e., decisions, may also be routed in physics (e.g., Schade and Sunder 2020).

that (singular universe) *determinism* – the basic premise of classical physics – does not matter for the free will debate, but clearly assuming that determinism might be correct – and opting for a compatibilist account.[2,3]

A somewhat special position has been introduced by Schade (2015) who not only justified the existence of free will via *quantum mechanics* (such as the above-mentioned libertarians; e.g., Kane 1985; Cogley 2015); but other than within libertarianism, an actual free will in the sense of being able to 'choose otherwise' was argued to be justifiable *without* indeterminism, within a framework offered by the multiverse interpretation of quantum mechanics, along with parallel times (for a more detailed treatment of the matter see the monograph by Schade 2018; see also Schade 2020). Other than in the Copenhagen (or 'standard') interpretation of quantum mechanics underlying libertarianism, there is no problem of randomness or 'luck' (see, e.g., the discussion in Cogley 2015) or other problems arising from an indeterminist notion of free will in a singular universe (see Schade 2020 and the discussion section of this paper) because the Schrödinger equation underlying the multiverse is *deterministic* (there are no random collapses of the wave function leading to indeterminism); and at the same time a *multiplicity of possibilities* exists given the *superposition* principle so that 'choosing otherwise' is in principle possible[4] in a more fundamental way than within compatibilist accounts.

Also, a set of arguments has been crafted within Schade (2015, 2018, 2020) suggesting that the multiverse might be the most *compelling* interpretation of quantum mechanics (see also the earlier, related thoughts on this by Menski 2000, 2005, 2007,

2. A more complete treatment of the many JCN-contributions on free will and their respective physical basis is beyond the scope of this paper.

3. It is important to note that the free will theorem in its strong form by Conway and Kochen (2009) shows that a couple of desirable conditions is inconsistent with determinism or, if two experimenters are free to choose certain measurements, then the outcomes of their measurements cannot be determined by the past. Let me note four things here. (1) It might not exactly be free will (in the philosophical sense) that this theorem is about, rather non-determinism, since the potential sources of freedom are not addressed. (2) The theorem clearly rejects classical physics in its idea of a clockwork universe. (3) Beyond the latter, it is not clear whether this theorem would imply any specific interpretation of quantum mechanics; indeed, one would expect that it rather does not, since it is often applauded for not leading to or requiring any specific theory of physics. (4) According to my view, past and future might be seen as problematic categories, if there is no linear flow of time (DeWitt 1967; Barbour 1999; Schade 2015 and 2018, chap. 3). For all those reasons, and although the theorem might suggest a clear "yes" to the answer raised in the title of this paper, this paper is *not* dealing in more detail with the free will theorem.

4. For more details/further conditions see the development in Schade 2018, chaps. 2, 3 and 4.

2010 or, e.g., by Deutsch 1997) and that there are good arguments existing in physics for parallel times (e.g., DeWitt 1967; Barbour 1999; Schade 2015, 2018, chap. 3).[5] Even though the line of arguments suggested by these authors – a review would be beyond the scope of this paper – might be seen as compelling by some, however, others might argue that the only way to convince them of that set of arguments would be *experimental*, and that only if at least the multiverse interpretation has been proven experimentally, they might perhaps consider the free-will argument presented by Schade (2015, 2018, 2020).

But can the multiverse be proven experimentally? This question has made its way into TV comedy, even. In *The Big Bang Theory*,[6] Dr. Sheldon Cooper asks his (only) prospective student Howard Wolowitz: "What is the correct interpretation of quantum mechanics"? And since Howard wants to be accepted as Sheldon's student, he answers: "As every interpretation gives exactly the same answer to every measurement they are all equally correct. However, I know you believe in the many-worlds interpretation, so I'll say that." This conversation states a well-known puzzle. In all standard physical experiments, we expect the same outcome no matter which interpretation is true, the standard or von Neumann (1932 [1996]) interpretation assuming a collapse of the wave function during measurement (leading to a singular reality but quantum indeterminism), or the many-worlds interpretation (Everett 1957), often called the quantum multiverse or theory of the universal wave function.[7]

In this paper, I am going to challenge the notion of the 'untestable multiverse,' even without presenting any experimental evidence. Instead, I am going to craft a type of experiment that is novel and that I suppose should actually be run, a type of experiment that might, in fact, discriminate between different interpretations of quantum mechanics. This type of experiment might be called 'psychophysical' (see Schade 2018; chaps. 12 and 13) and is based upon the understanding that there is a close connection between

5. Generally, the interpretation problem of quantum mechanics might be viewed as unsettled. The two most prominent interpretations are, indeed, collapse or 'standard' interpretation (von Neumann 1932 [1996]) and theory of the universal wave function (Everett 1957, building on Schrödinger 1926), also called 'many-worlds' or multiverse interpretation. However, the latter interpretation 'needs interpretation' (Albert and Loewer 1988). A novel proposal to this effect has been made by Schade (2018): the clustered-minds multiverse (CMM). The experiments suggested in this paper are imbued by the CMM and its top-down entanglement starting in consciousness, an idea refining Wigner's (1961) idea of measurement starting in consciousness and translating it to the multiverse, i.e., abandoning any collapse postulate (see also below).

6. Season 8, episode 2, "The Junior Professor Solution," CBS, Monday, September 22, 2014.

7. In this publication, I am only dealing with those two, not with other interpretations or modifications of quantum mechanics such as the Bohmian interpretation or the Penrose modification etc.

consciousness and free will on the one hand, and consciousness and the measurement problem on the other hand.[8] It is not psychophysical in the sense of, e.g., the Weber–Fechner laws of the relationship between physical stimuli and human perception. Instead, in the type of experiment that I am going to suggest, psychophysical is meant in a kind of reversed manner, as quantum optical experiments are supposed to be modified via specific manipulations of actual conscious observation.[9] The theoretical underpinning of the experiments is idealist in the sense of Plato (~ 360 B.C., 2000) or, based on quantum mechanics, Wigner (1961), but without running into any logical paradoxes (at least in conjunction with the idea of free will) as singular-universe approaches such as Wigner's typically do (Schade 2018, 2020 and the discussion section of this paper). The experiments will be idealist in a multiverse fashion as proposed within the *clustered-minds multiverse* (Schade 2018; for first thoughts on this, introducing the notion of weak and strong universes[10] as well as using the example of a torchlight[11], see also Schade 2015). At the same time, the type of experiment will be quite practical in nature, it can be implemented in a joint effort of experimental physicists and psychologists (or experimental economists, for that matter). It will also be fairly explicit and unambiguous, testing behavior of a quantum system against predictions based on a mathematical formula.

The contribution of the current paper is, thus, potentially large. It might form the basis of a new type of experiments that, in turn, help addressing the interpretation problem of quantum mechanics. It moreover, gives explicit directions as to how such experiments might practically be set up. It therefore helps advancing the issue as to how free will might be justified, if, (a) the line of arguments in Schade (2015, 2018, 2020) will be accepted as compelling and, within the proposed experiments, (b) the multiverse interpretation will turn out to be supported.[12] Some might be tempted to see

8. A novel type of experiments combining quantum mechanics and psychology has also been postulated by Mensky (2000; 2005).

9. A look at an exciting paper from Fedrizzi's work group (Proietti et al. 2019) shows that the general idea of observer dependence of quantum systems underlying also the experiments suggested in this paper is experimentally supported. However, the 'observers' in Proietti are photons, no actual, conscious observers.

10. This idea has been suggested to me in a discussion by Tanja Schade-Strohm.

11. This idea has been suggested to me in a discussion by Adam P. Taylor.

12. I am not dealing with the question, here, how parallel times might be tested. In a way, this question already has been positively tested (see the line of reasoning in Schade 2015 as well as 2018, chap. 3), even though some might debate this.

the type of experiments suggested, given their psychophysical nature and the test of an active role of consciousness, as a, potentially, direct proof of the existence of free will, circumventing the, perhaps 'inadmissible,' multiverse somehow, but in fact matters are fairly complicated and such a direct proof of free will is not feasible (see the discussion section).

The paper is organized as follows. In the next section, the most famous quantum optical experiment, the double-slit experiment and its most typical extension (eliciting so-called which-way information), will be introduced (including some historical information). The basic premises of the psychophysical experiments suggested in this paper are also outlined. The subsequent section specifies one of those psychophysical experiments in more detail, i.e., a double-slit experiment varying the *number of observers* as well as containing other treatments and provides a formula that is supposed to be tested within those experiments. The final section briefly extends the line of thoughts to related quantum-optical experiments that might be used for robustness checks of findings potentially generated at the double slit, dismisses singular-universe free will via the necessity of parallel times, and concludes.

Standard Quantum Optics versus 'Psychophysical' Experiments

The pre-version of the double-slit experiment is more than two hundred years old: Young let regular light pass through two parallel, vertical slits and demonstrated an interference pattern consistent with the wave nature of light. Thought experiments employing some fictitious versions of a double-slit experiment, typically having photons pass through the double slit and asking the question as to where the photon(s) passed through, i.e., so-called *which-way* experiments,[13] are about eighty years old. *Actual* double-slit experiments *with photons* are from the early sixties (Jönsson 1961). And smart *actual* which-way experiments have been carried out since more than thirty years (e.g., Mittelstaed et al. 1987; Menzel et al. 2012).

Whereas the behavior of particles[14] at the double slit, containing which-way information, has been interpreted by David Deutsch (1997, chapter 2) already as clear evidence in favor of the multiverse interpretation of quantum mechanics ('shadows'),

13. For more details see Feynman et al. (1965).

14. Some disagree that there are even particles and, instead, employ the notion of narrow wave packets (Zeh 2016).

others have disagreed. Indeed, the which-way thought experiments have originally been brought up within the paradigm of the Copenhagen interpretation of quantum mechanics. They have been an instrument of debate between Einstein, Heisenberg, Bohr etc. And although the novel findings in some actual which-way experiments at the double slit *seem* to undermine the complementary principle because they allow a high visibility of the interference pattern together with accurate which-way information (e.g., Menzel et al. 2012), an alternative view to David Deutsch's by researchers with a strong belief in the complementarity principle is justified by the fact that violations could always theoretically be justified somehow within the framework of the Copenhagen interpretation (even though the explanations gain more and more complexity). An example for this are theory and experiments on the *postselection principle* (Leach et al. 2014), to be discussed with a slightly different purpose below.

One thing has never been changed. There is only *one actual observer* in all those experiments.[15] And whereas the changes induced by such conscious observation have already had some argue that something 'special' is going on, that conscious observation is the final answer to the question as to where the 'Heisenberg cut' between the measurement and the 'to-be-measured' is supposed to be located, no one has actually 'played' with conscious observation in a more explicit way. That is the radical change in experimentation to be suggested in this paper. A psychophysical double-slit experiment is supposed to make measurements at the double slit with the number of observers and the type of feedback/information provided to them being experimentally manipulated.

It is *not* assumed that the manipulation of conscious observation will change the frequencies of photons flying through one slit or the other; but instead it is assumed that the *tension* between the strength of the interference pattern on the one hand and the precision of information regarding the 'which-way' question can be manipulated via the type of conscious observation. This is captured within a formula, to be introduced in the next section, against which the experimental treatments are testing.

The type of experiments to be suggested is also inspired by a set of experiments published by Radin et al. (2012). Radin et al. demonstrate that asking individuals to direct their *attention* more or less to the double slit has an influence on the strength of the observed *interference pattern*. Or in other words, these authors' findings indeed suggest that it is *consciousness* that matters with respect to the results that are observed at the double slit. However, their results do not serve to answer the question as to which

15. In Proietti et al. (2019) there seems to be more than one observer, but only the final observer, the experimental scientist, is 'real,' the others are photons, in the role of 'observers.'

interpretation of quantum mechanics might be correct. And therefore, following the set of arguments as well as the literature presented in the introduction, their results do not allow to answer the question whether physics is consistent with free will or not.

Also, whereas the approach to be introduced below in this article might be labelled psychophysical, Radin et al.'s (2012) approach might rather be labelled 'para-psychophysical'[16] as will be shown. Anyway, in the experiments carried out by Radin et al. (2012, 159), the dependent variable that was supposed to reflect changes due to the experimental treatments was defined in the following way:

> To measure perturbations in the wavefunction, the interference pattern recorded by the line camera was analyzed with a fast Fourier transform to quantify the power associated with the two dominant spatial wave-lengths: a shorter wavelength associated with the double-slit interference pattern (call this power P_D) and a longer wavelength associated with the diffraction pattern produced by each slit (P_S) (…). The fraction of (log)spectral power associated with the interference pattern was $D=[P_D/(P_D+P_S)]$, and that with the diffraction pattern was $S=[P_S/(P_D+P_S)]$. The ratio of these fractions, $R=D/S$, was the preplanned variable of interest.

The definition of those variables is reported in detail here for the sake of comparison with formula (1) that I am below suggesting should be tested to support (an idealist version of) the multiverse interpretation of quantum mechanics.[17] It will also be a test integrating the interference pattern, but in a different way and in conjunction with other information (see the next section).

Coming back to the para-psychology label that I have used above for Radin et al.'s work, this might be evidenced by the following quote, enriched by insertions in italics by myself (Radin et al. 2012, 160):

> The consciousness collapse hypothesis predicted that the act of focusing attention toward the double-slit *without any direct connection*

16. One might also call those experiments psychokinetic. Although effects of consciousness on the quantum might be considered part of 'normal' physics, depending on the interpretation of quantum mechanics chosen, the mechanism that Radin et al. (2012) are focusing on is psychokinetic because only attention of consciousness directed at the double slit (and no direct observation of the quantum system) is analyzed.

17. Such a proof will be *general* with respect to the multiverse interpretation, but it will also show that an idealist version of it will be the appropriate framework.

> *between observer and quantum system* would cause *R (the spectral-ratio value)* recorded during attention-toward epochs to decrease as compared to during attention-away epochs.

In contrast, the role of the observer that I am proposing to be analyzed will turn out to be 'conservative,' bearing the potential of a higher acceptability by mainstream physics (see the next section). Naturally, given the somewhat para-psychological nature of Radin et al.'s experiments, they have been criticized by some realists on various grounds. Also quite naturally, Radin et al. have refuted their criticism. It is beyond the scope of this study to report on this discussion in detail; also, as said, the type of observation suggested in the experiments propsed here will be more traditional in nature, so that the criticism and its discussion are simply not relevant here.

Specifying a Double-slit Experiment with Dual Observers and Other Experimental Conditions

Basic Premises

If one starts with measurement – i.e., finally *locates* the measurement problem – in consciousness, one gets to the concept of top-down decoherence (see Schade 2018, chap. 2; see also Bacciagaluppi 2020), or, more neutral, to *top-down entanglement*.[18] What is meant with this is that (the in principle) non-directional quantum correlations actually *start* in consciousness, so that consciousness is not the end of the chain of quantum correlations (some epiphenomenon upon the workings of physics) but its source. With top-down entanglement in a multiverse, quantum systems with two real, concurrent observers might, especially under conditions of diverging goals, differ in their physical consequences from single-observer systems (note that this idea is somewhat related to that of Wigner's friend (Wigner 1961; Proietti et al. 2019), but within a different interpretation of quantum mechanics and using a different implementation). Note that

18. 'Neutrality' means absence of operations leading to a *reduced density matrix*. This operation, typically the second stage in decoherence analysis, is sometimes 'accused' to introduce collapse through the 'backdoor,' and would hence not be 'neutral' anymore with respect to the interpretation problem of quantum mechanics. For a critical discussion of not just staying with quantum correlations (entanglement) but also calculating a reduced density matrix see Zeh (2012, 77–84); see also Schade (2018, chap. 2).

showing this divergence would be a *general* proof of the multiverse, not only a special version of it.

Proposing the Type of Experiment To Be Run

Let me now describe a version of the experiment that I propose should be run, based on the double-slit experiment. The basic setup is as usual, photons are fired at some plate, pierced in the form of two parallel slits (with a certain, critical distance), and there is a measurement of where the photons pass through: left slit *or* right slit?, as well as of the interference pattern at some screen, located in appropriate distance from the plate. Within that framework, however, and this is the novelty, the following six treatments (= experimental conditions) are to be implemented:

Treatment basic (*b*) will be run without anyone observing which-way observation, treatment *so* with single observers and rewards, coupled to one of two outcomes of the quantum experiment, i.e., whether the photon passes through the right or the left slit. Another four treatments will implement two concurrent observers (i.e., observer pairs) that both observe the *same* quantum experiment. This part of the experiment will implement a 2 x 2 design with the first factor (preferences) involving the two steps: (1) aligned (*do-ai*) versus (2) conflicting rewards (*do-di*), i.e., observers getting rewards for the same or for different outcomes of the quantum experiment. The second factor (information) will implement the following two steps: (i) no information on the outcomes that the other has observed, and (ii) information on the observed outcomes by the other player; in light of the below discussion on postselection and given the fact that postselection might be seen as equivalent to communication, the interpretation of that factor, however, will have to be carefully pondered.[19]

Experimental Treatments and Mathematical Ordering

The theoretical idea behind those manipulations is *not* to find any differences in the frequency of the two outcomes (i.e., the photons passing more through either the right slit or the left slit) of the respective quantum-optical experiments, e.g., depending on

19. Given the discussion in the next but the following subsection, the most interesting results regarding the factor information might materialize in the form of *interaction effects* with the factor preferences. It is beyond the scope of this contribution to explore this in more detail.

some characteristics of the observers – this would be a parapsychological prediction – but, as already mentioned, to put the quantum system under more or less 'stress.'

What is exactly meant with putting the quantum system under 'stress?' In the clustered-minds multiverse, overall consciousness (i.e., the total of consciousness across all versions of an individual) will have to 'split attention' between different realities, in our case 'marked' by different passing of the photon through either the left or the right slit.[20] But since rewards are coupled to either the *same* observation (left slit *or* right slit) being preferred by *both* observers or to *two different ones*, realities with more or less tension regarding consciousness to be allocated between realities will emerge (see, for the basis of this idea, Schade 2018, chap. 8, page 139-141, especially formulas 8.1 and 8.2). What are the implications of these thoughts? High fringe visibility (interference) is expected in treatment *b*; in the *so* treatment, the sum of V^2 (visibility of the interference, squared) and P^2 (which-alternative information, squared) should be similar to the respective sum in *b* (duality principle in its up-to-date form: e.g., Greenberger and Yasin 1988). If there are two observers with the same goals, the stress on the system should be larger than with singular observers, but smaller than with conflicting goals; i.e., regarding the latter, the 'stress' put on the quantum system to keep an intact interference structure on the one hand (e.g., quite figural, 'shadows of other realities' in David Deutsch's not undebated view; Deutsch 1997, chap. 2, see above), but 'having to provide' two different realities to the two observers on the other hand, might considerably enhance the sum of V^2 and P^2. For the sake of brevity and to reduce complexity, the second factor (information provision) will not be discussed here in more detail (and might, anyway, require a deeper analysis; see the next subsection) and will thus not be integrated into the following, preliminary formula. (Think of it, for now, as this factor being fixed at "no information on the outcomes of the other player" for formula (1)). It will also be left open here whether (in any of the treatments) $V^2 + P^2$ might become larger than one or not (typically: = 1, for pure states, ‹ 1, for mixed states, but also › 1 under special conditions; see the discussion in the next subsection) (e.g., Leach et al. 2016). Given those simplifications, the suggested experiment is going to test the following set of conditions:

$$\mathcal{V}_b^2 + \mathcal{P}_b^2 \approx \mathcal{V}_{so}^2 + \mathcal{P}_{so}^2 < \mathcal{V}_{do-ai}^2 + \mathcal{P}_{do-ai}^2 \ll \mathcal{V}_{do-di}^2 + \mathcal{P}_{do-di}^2 \qquad (1)$$

20. It should be noted that observing a photon flying through either the right slit or the left slit in a quantum apparatus is sufficient to generate two different realities. Everett (1957) would associate this event with a splitting of the universe.

Avoiding the Relevance of Alternative Explanations: Postselection and the Like

In a paper by Menzel et al. (2012), clear interference fringes are observed together with sharp which-way information. Leach et al. (2016) discuss (and experimentally show) as to *how* the sum of V^2 (visibility of the interference, squared) and P^2 (which-alternative information, squared) may reach values up to two, specifically (3):

$$\mathcal{V}_{\hat{\pi}1}^2 + \mathcal{P}_{\hat{\pi}2}^2 \leq 2 \qquad (2)$$

Surpassing the sum of 1 may be feasible when "measuring visibility and predictability are conditioned on different postselections $\hat{\pi}1$ and $\hat{\pi}2$" (4), with the most extreme case occurring when the two are orthogonal to each other, then leading to a sum of 2.[21] So potentially, such effects could be alternative explanations for the results to be expected in the proposed experiment, to claiming the effects of a manipulation of the number of observers etc. (see above) and thus a potential, alternative explanation to the one suggested: the existence of a quantum multiverse. So this alternative explanation for the experimental results would be a serious threat to the theoretical development suggested in this paper, including the idea of constructing the basis for the existence of an actual free will. Even more plausible, and as already mentioned, they could be an explanation for the potential effect of exchanging information on the observed state of the system between the dual observers (in some of the treatments proposed above). Indeed, as Leach et al. (2016, 5) note,

> We show that if a qubit is coupled to its environment, it becomes possible to obtain simultaneous high values for conditional measures of visibility and predictability. (…) We note that although our experimental procedure allowed us to *purposely* obtain simultaneous high values which lead to an obvious violation of an algebraic bound, there can be realistic experimental cases where an *inadvertent* postselection could be performed without necessarily obtaining a clear violation. In these cases, detecting the loophole might be much more difficult. [Italicizing by the author of the current article]

21. A structurally similar approach, requiring the same ingredients and leading to qualitatively comparable outcomes is direct measurement relying on weak values (see Lundeen et al., 2011; Salvail et al., 2013).

When running the test of the multiverse suggested above, one thus has to avoid this problem, keeping an eye on *inadvertent cases* of postselection. One safeguard arises from avoiding experimental manipulations implementing *any* apparent changes of the physical environment relevant within the quantum-mechanical calculus (this is fully avoided, I insist, in the variation of the number of observers suggested in the current paper). Another safeguard, most naturally, is taking changes explicitly into account such as the change of information regime in some of the experimental treatments and *modelling* them in terms of postselection.

Discussion, Further Steps, and Conclusions

Running the proposed experiment and avoiding (or explicitly modelling) postselection and other potential threats of internal validity potentially leads to a 'proof' of the multiverse and this, in turn, to a theoretical basis for an actual free will. Let me assume that the results are perfectly in tune with the conditions specified in formula (1) (leaving out, once more, the information conditions or assuming that they, indeed, be explicitly modelled). Quite naturally, nevertheless, such farfetched consequences would not be applauded unless some robustness checks, some related experiments have been run. One might, therefore, not only carry out double-slit experiments but also other quantum-optical experiments, e.g., *interferometer* experiments; implementing the same *type* of treatments described above, but with technical modifications appropriate for the different optical setup at hand (details are beyond the scope of this contribution).

Moreover, I have mentioned that some might view the type of experiments suggested as a direct proof of free will (if they generate the predicted results), perhaps hoping to 'circumvent' the multiverse, somehow. Besides the fact that those experimental results would, anyway, *be* a proof of the multiverse, seeing them as a direct proof of free will abstracts from some complexities that have to be taken into account. Just a few words on this. Let me assume that two observers, depending on whether their preferences are aligned or divergent, are indeed able to produce more or less tension within a quantum system. Then this will be interpreted as a proof of the multiverse and will generate the planned, indirect proof of free will, if, in addition, the framework presented in Schade (2015; 2018; 2020) is seen as compelling.

But clearly, the experimental results would tell us more: a story of an idealist version of the world (*Maja*, in Indian philosophy), of 'mind over matter,' without (other than in Radin et al. 2012) any necessary recurrence to parapsychology and with real observers

rather than photons (other than in Proietti et al. 2012). And, if it is accepted that mind is ruling matter rather than the other way around, wouldn't free will arise as a natural consequence? Moreover, wouldn't the active role of consciousness proven within such experiments hint in a similar directions as the implications derived from the strong free-will theorem (Conway and Kochen 2009)? According to that version of the free-will theorem, *if* decisions *what* to measure can freely be implemented by observers, the future is non-determined by the past. But there are a few problem with this, part of them already specified in footnote 3. One of them was not mentioned there: Our observers are *not* free to choose what to measure. Another had been mentioned: At least the *notion of a regular flow of time* (past and future being integral parts of the theorem) is quite problematic in the context of free will (see below). So at least, and as already conjectured in footnote 3, the free-will theorem is not terribly helpful, here.

But still, the temptation to directly 'leap' to free will, not taking any 'detour' via the multiverse and a somewhat complex theoretical development, might be large. So, again, why is the multiverse so important to justify free will, why couldn't we just stay with the comfortable, well-known idea of a singular universe? The reason is that singular-universe free will may not be seen as even possible. Let me reference my own work, here (Schade 2020, 324-325) (insertions in brackets are added within this contribution):

> Many changes in the weltanschauung [compared to classical, deterministic physics] are already realized within the standard, singular-universe interpretation of quantum mechanics (collapse theory/reduction postulate). And some researchers indeed use this interpretation as a basis of free will (e.g., Kane 1985; Stapp 2017; Laskey 2018). Of the singular-universe approaches, I regard Laskey's (2018), built on Stapp (who points for inspiration to von Neumann), to be the most creative and advanced. According to this theory, free will is related to the choice of what to measure and to the quantum Zeno effect (Misra and Sudarshan 1977). The anti-free will evidence presented by Libet and coauthors (e.g., Libet et al. 1982) as well as his neuroscience followers (e.g., Soon et al., 2008) is my main reason for suggesting a novel, multiverse-based alternative to the various collapse/singular-universe versions of free will.

Specifically, in all the experiments by Libet and coauthors as well as in Soon et al. *consciousness is running after the fact*, apparently an obstacle for most individuals' common-sense idea of free will somehow being related to some choices made in

consciousness and then 'executed' (note that many sophisticated versions of free will do not differ much from this notion, whereas, indeed, matters are getting slightly less straightforward in the multiverse). Interestingly, this obstacle can be circumvented, but, in my opinion, *only* in the multiverse, not within any singular-universe theories. Here is a fairly condensed version of the argument that could be made (Schade 2020, 325; see also Schade 2015):

> (…) singular-universe quantum theories, with their implied irreversibility of actions, cannot rule out the inference that free will is a mere illusion. (…) ruling out no-free will inference from the Libet et al. data, one needs parallel times, or times as special cases of parallel universes (…). The basis for this, in turn, is provided within the Wheeler/ DeWitt equation (DeWitt 1967) linking general relativity and the Schrödinger equation, where time as a variable disappears. In the same way in which the—unaltered and unaccompanied—Schrödinger equation is a multiverse equation, the Wheeler/ DeWitt equation also is. Thus, it is the multiverse perspective that rules out the Libet evidence against free will.

So is it really necessary to take a view into cosmology and cosmological equations (such as the Wheeler/DeWitt equation) and into the problem of time (a philosophical term crafted to address the problem that the time variable disappears in the Wheeler/DeWitt equation), into the multiverse, anyway, to be able to justify free will? And are we really prompted to run the experiments suggested in this article to first prove the existence of a quantum multiverse and then, using the line of arguments suggested in Schade (2015, 2018, 2020), are able to provide a framework, the clustered-minds multiverse, that accommodates for free will?

Well, a pure theorist could perhaps save on the experiments suggested here, and just buy into the free-will arguments crafted in connection with the clustered-minds multiverse. Whereas I am personally sympathetic to this position, because with respect to free will and the multiverse, theories from different areas fit together like the pieces of a puzzle (see Figure 1 in Schade 2015), I also clearly understand that some would require an empirical proof, and I would personally love to be able to deliver one.[22] Therefore, I

22. Actually, when I am not writing about quantum decision making and free will, I do run many laboratory experiments on decision making or analyze large datasets and publish the results in Journals on Economic Psychology, so that I need not to be convinced of the beauty of experimentation.

suppose that for most people the answer to the above questions would be: "yes", and I very much hope that the suggested experiments will finally be run.

References

Albert, David, and Barry Loewer. 1988. "Interpreting the Many Worlds Interpretation." *Synthese* 77: 195–213.

Barbour, Julian. 1999. *The End of Time: The Next Revolution in our Understanding of the Universe*. UK: Weidenfeld & Nicholson.

Cogley, Zac. 2015. "Rolling Back the Luck Problem for Libertarianism." *Journal of Cognition and Neuroethics* 3 (1): 121–137.

Conway, John H., and Simon Kochen. 2009. "The Strong Free Will Theorem." *Notices of the AMS* 56 (2): 226–232.

DeWitt, Bryce S. 1967. "Quantum Theory and Gravity. I. The Canonical Theory." *Physical Review* 160: 1113–1148.

Deutsch, David. 1997. *The Fabric of Reality: Towards a Theory of Everything*. Middlesex: Penguin Books Ltd.

Everett, Hugh, III. 1957. "'Relative State' Formulation of Quantum Mechanics." *Reviews of Modern Physics* 29: 454–462.

Feynman, Richard P., Robert B, Leighton and Matthew Sands. 1965. *The Feynman Lectures on Physics*, Vol. 3, 1.1-1.8. Reading, Mass: Addison-Wesley.

Greenberger, Daniel M. and Allaine Yasin. 1988. "Simultaneous Wave and Particle Knowledge in a Neutron Interferometer." *Physics Letters A* 128 (8): 391-394.

Jönsson, Claus. 1961. "Elektroneninterferenzen an mehreren künstlich hergestellten Feinspalten." *Zeitschrift für Physik* 161 (4): 454–474.

Kane, Robert H. 1985. *Free Will and Values*. Albany, NY: State University of New York Press.

Laskey, Kathryn B. 2018. "Acting in the World: a Physical Model of Free Choice." *Journal of Cognitive Science* 19 (2): 125–163.

Leach, Jonathan, Eliot Bolduc, Filippo M. Miatto, Kevin Piché, Gerd Leuchs, and Robert W. Boyd. 2016. "The Duality Principle in the Presence of Postselection." *Scientific Reports* 6 (1): 19944. https://doi.org/10.1038/srep19944.

Libet, Benjamin, Elwood W. Wright, Jr., and Curtis A. Gleason. 1982. "Readinesspotentials Preceding Unrestricted 'Spontaneous' vs. Pre-Planned Voluntary Acts." *Electroencephalography and Clinical Neurophysiology* 54: 322–335.

Lundeen, Jeff S., Brandon Sutherland, Aabid Patel, Corey Stewart, Charles Bamber. 2011. "Direct Measurement of the Quantum Wavefunction." *Nature* 474: 188-191.

Mensky, Michael. B. 2000. "Quantum Mechanics: New Experiments, New Applications, and New Formulations of Old Questions." *Physics – Uspekhi* 43: 585-600.

Mensky, Michael. B. 2005. "Concept of Consciousness in the Context of Quantum Mechanics." *Physics – Uspekhi* 48: 389-409.

Mensky, Michael. B. 2007a. "Quantum Measurements, the Phenomenon of Life, and Time Arrow: The Great Problems of Physics (in Ginzburg's Terminology) and Their Interrelation." *Physics – Uspekhi* 50: 397–407.

Mensky, Michael. B. 2010. *Consciousness and Quantum Mechanics: Life in Parallel Worlds*. Singapore: World Scientific Publishing Co.

Menzel, Ralf, Dirk Puhlmann, Axel Heuer and Wolfgang P. Schleich. 2012. "Wave-Particle Dualism and Complementarity Unraveled by a Different Mode." *Proceedings of the National Academy of Sciences* 109: 9314–9319. doi: 10.1073/pnas.1201271109.

Misra B. and E.C.G. Sudarshan. 1977. "The Zeno's Paradox in Quantum Theory." *Journal of Mathematical Physics* 18: 756–763.

Mittelstaedt, Peter, A. Prieur and R. Schieder. 1987. "Unsharp Particle-wave Duality in a Photon Split-beam Experiment." *Foundations of Physics* 17 (9): 891–903.

Neumann, Johann von. (1932) 1996. *Mathematische Grundlagen der Quantenmechanik*. 2nd ed. Berlin: Springer-Verlag.

Plato. (* 360 B.C.) 2000. "Timaeus." Translated by Donald J. Zeyl. Indianapolis, IN: Hackett Publishing Company, Inc.

Proietti, Massimiliano, Alexander Pickston, Francesco Graffitti, Peter Barrow, Dmytro Kundys, Cyril Branciard, Martin Ringbauer, and Alessandro Fedrizzi. 2019. "Experimental rejection of observer-independence in the quantum world." arXiv:1902.05080v1.

Radin, Dean, Leena Michel, Karla Galdamez, Paul Wendland, Robert Rickenbach and Amaud Delorme. 2012. "Consciousness and the double-slit interference pattern: six experiments." *Physics Essays* 25: 157-171.

Salvail, Jeff Z., Megan Agnew, Allan S. Johnson, Eliot Bolduc, Jonathan Leach, and Robert W. Boyd. 2013. "Full Characterization of Polarization States of Light via Direct Measurement." *Nature Photonics* 7: 316-321.

Schade, Christian D. 2015. "Collecting Evidence for the Permanent Coexistence of Parallel Realities: An Interdisciplinary Approach." *Journal of Cognition and Neuroethics* 3 (1): 327–362.

Schade, Christian D. 2018. *Free Will and Consciousness in the Multiverse: Physics, Philosophy and Quantum Decision Making*. London: Springer.

Schade, Christian D. 2020. "Free Will in the Clustered-minds Multiverse, and Some Comments on S. Sarasvathy's 'Choice Matters'." *Mind and Society* 19: 323-330.

Schade, Christian D., and Shyam Sunder. 2020. "Physics and Decisions: An Exploration." *Mind and Society* 19: 287-292.

Schrödinger, Erwin. 1926. "Quantisierung als Eigenwertproblem." *Annalen der Physik* 79: 361–376.

Soon, Chun S., Marcel Brass, Hans-Jochen Heinze, and John-Dylan Haynes. 2008. "Unconscious Determinants of Free Decisions in the Human Brain." *Nature Neuroscience* 11: 543–554.

Stapp, Henry P. 2009. *Mind, Matter, and Quantum Mechanics*. 3rd ed. London: Springer.

Vihvelin, Kadri. 2015. "How Not To Think about Free Will." *Journal of Cognition and Neuroethics* 3 (1) : 393–403.

Wigner, Eugene P. 1983 [1961]. Remarks on the mind-body question. Reprinted in *Quantum Theory and Measurement*, ed. by John A. Wheeler and Wojciech H. Zurek, 285-302. Princeton: Princeton University Press. [Original German title: Die gegenwärtige Situation in der Quantenmechanik.]

Zeh, H. Dieter. 2012. *Physik ohne Realität: Tiefsinn oder Wahnsinn?* London: Springer.

Zeh, H. Dieter. 2016. "The Strange (Hi)story of Particles and Waves." *Zeitschrift fuŕ Naturforschung A* 71: 195–212.

cognethic.org