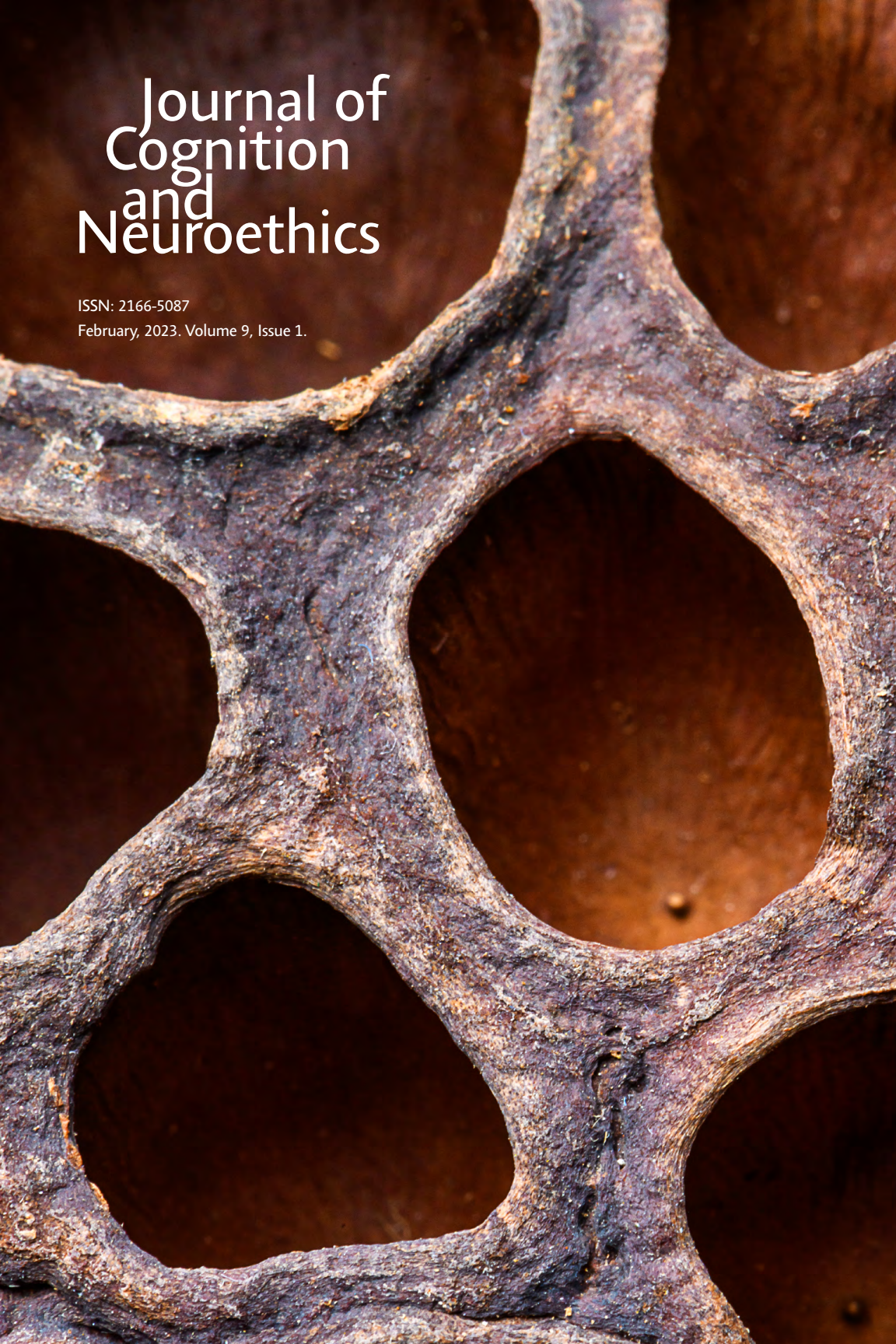

Journal of Cognition and Neuroethics

ISSN: 2166-5087
February, 2023. Volume 9, Issue 1.

Journal of Cognition and Neuroethics

Managing Editor
Jami L. Anderson

Production Editor
Zea Miller

Publication Details
Volume 9, Issue 1 was digitally published in February of 2023 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

The Neuroethics of Memory's Social Value: To What Extent Can Neurotechnologies That Manipulate Memory Be Permitted?

Eisuke Nakazawa
The University of Tokyo

Koji Tachibana
Chiba University

Keiichiro Yamamoto
National Center for Global Health and Medicine

Akira Akabayashi
The University of Tokyo
New York University

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2023. Volume 9, Issue 1.

Citation
Nakazawa, Eisuke, Koji Tachibana, Keiichiro Yamamoto, and Akira Akabayashi. 2023. "The Neuroethics of Memory's Social Value: To What Extent Can Neurotechnologies That Manipulate Memory Be Permitted?" *Journal of Cognition and Neuroethics* 9 (1): 1–11.

Eisuke Nakazawa

Eisuke Nakazawa (ORCID 0000-0002-3320-3811), PhD, is a lecturer of Biomedical Ethics at School of Public Health, Faculty of Medicine, The University of Tokyo, Japan. He received his doctorate in in Philosophy of Science from the University of Tokyo. His current research area is Biomedical Ethics, Neuroethics and Philosophy of Science. His recent publications include Sadato N, Morita K, Kasai K, Fukushi T, Nakamura K, Nakazawa E, Okano H, Okabe S. 2019. Neuroethical issues of the Brain/MINDS project of Japan. *Neuron* 101 (February 6, 2019):385–389. https://doi.org/10.1016/j.neuron.2019.01.006. And Ino Y, Nakazawa E, Akabayashi A. 2019. Health and welfare in Japan. *The Lancet* 394 (10209):1614–1615. (November 2, 2019) https://doi.org/10.1016/S0140-6736(19)31805-7. He is a member of the International Neuroethics Society.

Koji Tachibana

Koji Tachibana (ORCID 0000-0001-5203-7081), PhD, is an assistant professor of philosophy, Faculty of Humanities, Chiba University, Japan. He received his doctorate in Philosophy of Science from the University of Tokyo. His current research area is Aristotle's Ethics, Contemporary Virtue Ethics, Neuroethics and Space Ethics. His recent publications include Konrad Szocik, Mark Shelhamer, Martin Braddock, Francis A. Cucinotta, Chris Impey, Pete Worden, Ted Peters, Milan M. Ćirković, Kelly C. Smith, Koji Tachibana, Michael J. Reiss, Ziba Norman, Arvin M. Gouw, Gonzalo Munévar. 2021. Future Space Missions and Human Enhancement: Medical and Ethical Challenges. *Futures* 133 (October 2021) https://doi.org/10.1016/j.futures.2021.102819. And Koji Tachibana. 2019. Nonadmirable moral exemplars and virtue development. *Journal of Moral Education* 48(3) 346-357 (July 2019). https://doi.org/10.1080/03057240.2019.1577723. He is a member of the Aristotelian Society.

Keiichiro Yamamoto

Keiichiro Yamamoto (ORCID 0000-0002-4763-4030), PhD, is head of the Office of Bioethics at the National Center for Global Health and Medicine in Tokyo, Japan. He received his doctorate in Ethics from the Kyoto University. His current research interests lie in Moral Philosophy, Bioethics, Research Ethics . His recent publications include Eisuke Nakazawa, Keiichiro Yamamoto, Alex John London, Akira Akabayashi, Solitary death and new lifestyles during and after COVID-19: wearable devices and public health ethics, *BMC Medical Ethics* 22 (89) (July 2021) https://doi.org/10.1186/s12910-021-00657-9; Kenji Matsui, Keiichiro Yamamoto, Shimon Tashiro, Tomohide Ibuki, A systematic approach to the disclosure of genomic findings in clinical practice and research: a proposed framework with colored matrix and decision-making pathways, *BMC Medical Ethics* 22 (168) (December 2021) https://doi.org/10.1186/s12910-021-00738-9. He is a member of the American Society for Bioethics and Humanities.

Akira Akabayashi

Akira Akabayashi (ORCID 0000-0003-0811-1955), MD, PhD, is Professor of Biomedical Ethics at School of Public Health, Faculty of Medicine, The University of Tokyo, Japan, and Adjunct Professor of Medical Ethics at Division of Medical Ethics, New York University School of Medicine, New York, USA. His research interests span cross-cultural bioethics, global bioethics, medical/clinical ethics such as informed consent, organ transplantation, and end-of-life issues, public health ethics, research ethics, and bioethics policy making. As an academic researcher, he has published more than 180 original articles and more than 20 books or chapters in English in addition to many Japanese publications. He was a former member of the board of directors of International Association of Bioethics. He was also honored as a *Fellow of The Hastings Center* (USA) in 2008. He is currently an editorial board of *Journal of Medical Ethics, Cambridge Quarterly of Healthcare Ethics*, *BMC Medical Ethics, and Asian Bioethics Review.*

Corresponding Author
Akira Akabayashi, MD, PhD
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
Department of Biomedical Ethics, University of Tokyo Faculty of Medicine
akira.akabayashi@gmail.com; akirasan-tky@umin.ac.jp

Funding
This paper is supported by JST RISTEX Grant Number JPMJRX21XX, and SPS KAKENHI Grant Number JP20H05717.

Statements and Declarations
The authors have no financial or non-financial conflicts of interest or competing interests to disclose.

The Neuroethics of Memory's Social Value: To What Extent Can Neurotechnologies That Manipulate Memory Be Permitted?

Eisuke Nakazawa, Koji Tachibana, Keiichiro Yamamoto, and Akira Akabayashi

Abstract

Memory manipulation technology has the potential to disrupt the social value of memory. Although the development of pharmacological interventions aimed at alleviating the fear conditioning that causes post-traumatic stress disorder has been challenging, there has now been significant technical progress in manipulating fear conditioning through the use of neurofeedback technologies. The manipulation of memory is often criticized on the basis of the social value of memory, as well as on the grounds of preservation of personal identity and authenticity. The social value of memory has been overestimated and has normatively hindered the application of memory manipulation technologies. Even if the manipulation of memory interfered with the social value of memory, it would be permissible, provided it was limited. When a memory is related to an illegal or unethical behavior, the preservation of memory's social value should be given priority. The modification of collective memory can also lead to historical revisionism. Recognizing the social nature of memory, and endorsing, preserving, and narrating it, is supererogatory.

Keywords

Memory, Social Value of Memory, Neurofeedback, PTSD

1. Introduction

Memory manipulation technology holds promise as a treatment for post-traumatic stress disorder (PTSD). Treatment of PTSD involves both symptomatic treatment—the administration of drugs for depression and anxiety related to traumatic memories—and cognitive behavioral therapy (CBT) aimed at controlling the traumatic memories that are the root cause (i.e., alleviating fear conditioning). As for the latter, cognitive processing therapy, cognitive therapy, eye movement desensitization and reprocessing, individual CBT with a trauma focus (undifferentiated), and prolonged exposure have been strongly recommended by the International Society for Traumatic Stress Studies (ISTSS) (Bisson et al. 2019).

In 2018, the ISTSS also proposed neurofeedback and *yoga* as treatments for PTSD based on emerging evidence, while adding a warning that as of yet the quality of the evidence is low (Bisson et al. 2019). Neurofeedback has also been used to alleviate fear

conditioning in traumatic memory, including decoded neurofeedback studies using fMRI (Koizumi et al. 2016) as well as research using electroencephalograms (van der Kolk et al. 2016; Steingrimsson et al. 2020). With regard to pharmacological interventions aimed at alleviating the fear conditioning that causes PTSD, a number of randomized controlled tests of drugs such as hydrocortisone and propranolol have been conducted, but with no major success (Astill Wright et al. 2019). Yet, studies using animal models have given us hope for the future prospects of research on memantine and other drugs (Ishikawa et al. 2019).

Given the clinical benefits of controlling traumatic memories and alleviating fear conditioning, we examine in this study the social implications of memory manipulation, with a particular focus on the social value of episodic memory (Liao and Sandberg 2008). Episodic memory is the privileged possession of individuals who can recall it. Considering the interiority of such memories, it may sound strange to insist that memories have social value. And yet, episodic memory includes descriptions of events that have occurred in this world where we have lived with others. Being conveyed to other people through language, episodic memory can be informed to others and develop intimacy and empathy with them (Alea and Bluck 2003). Thus, episodic memory has, at its core, social aspect (Alea and Bluck 2003). A representative type of episodic memory with social value is eyewitness memory, which frequently plays an important role in criminal trials (Lacy and Stark 2013). One can easily imagine how people's fates could turn on whether or not such memories exist. If the results of a specific incident have social impact, then it naturally follows that an eyewitness' memory of the incident has social value.

The social value of episodic memory is produced by primarily being mediated through language. By definition, episodic memory, which is episodic, includes both cognitive and non-cognitive aspects (Tulving 1972). Cognitive aspects can be communicated to others through language. Non-cognitive aspects, on the other hand, are the emotions and impressions connected to a particular episodic memory. Because manipulating technology of an individual's episodic memory usually intervenes the non-cognitive contents of memory, it may be tempting to conclude that it has no impact on the social value of memory. Alleviating fear conditioning involves the manipulation of the non-cognitive contents of episodic memory. The social value of memory must be communicable and comprehensible through language, and if we hold that it is possible to consolidate only the cognitive contents of episodic memory, it is understandable that we may be tempted by such conclusions. However, it appears to be conceptually and empirically impossible to regard the cognitive and non-cognitive contents of episodic memory as two parts that have no influence on one another. Moreover, it may be possible

that, by sensing the demeanor of an individual who is recalling an episodic memory, non-cognitive contents can be directly shared through compassion. Therefore, manipulating episodic memory, regardless of what form the manipulation takes, has the potential to affect the social value of memory.

Several researchers have discussed the social value of memory and its dampening (Liao and Sandberg 2008; Kolber 2006). These pioneering studies show that the social nature of memory is sharply opposed to memory manipulation. Liao and Sandberg (2008), for example, refer to the memory of Neil Armstrong, who was the first to reach the moon, and to the memory of Holocaust victims. Kolber (2006) goes further and argues that the need for medical treatment can defend against memory dampening. Although there are many other issues to be discussed regarding memory manipulation, such as personal identity and authenticity (Lavazza 2018; Lavazza 2019), we will focus on the social value of memory in this paper and pursue it in depth. Thus, we address the following questions in the present study. If a particular episodic memory possesses social value, how should we ethically evaluate medical interventions that erase or alter it? Is preserving such memories as they are our duty, or is it an act of supererogation?

2. Discussion

2.1. PTSD Treatment and Memory Manipulation

The disruption of episodic memory with social value during the course of treating PTSD, regardless of the type of disruption, must be permitted at the social level. It fits within the patient's right to treatment. For patients who possess traumatic memories and suffer from PTSD symptoms, to demand the preservation of such memories would place an additional burden on those deeply hurt by past events. This should be avoided. Consequently, even if a memory that possesses social value should disappear from existence by manipulation, society must simply accept it.

Taking into account the above discussion, let us consider the following scenario. The episodic memory possessed by Person A is the cause of A's PTSD. However, the health and well-being of someone else, Person B, depends on that same episodic memory. Imagine that A is an eyewitness in a murder case, and B is a family member of the victim. The episodic memory that threatens A's mental state is for B a priceless last memory of that family member, and its loss represents the disappearance of something very important to B. Even in this particular scenario, demanding the patient suffering from PTSD symptoms

to preserve the memory is impermissible. This is because, while A's episodic memory directly and causally affects his or her health, the causal relationship between A's episodic memory and B's health is ad hoc. Consider also the following extreme hypothetical. Person C is a witness to a massacre that occurred during a 20th century war, and now C is the only surviving witness. The preservation of C's memory has value to society, and C's memory is proof that these many victims existed and were killed during the war. Nonetheless, demanding C to stop PTSD therapy to preserve that memory would not be justifiable because such prohibition must require C to be the victim for society and such requirement is against Kantian notion of moral. If society were to force that C preserve the memory that cause PTSD, C would be regarded not as a person from the society, but as a mere storage device, i.e., something instrumental.

2.2. Boundary Between Treatment and Enhancement

The use of neurotechnology to manipulate episodic memory that goes beyond PTSD treatment requires a more careful ethical analysis. Techniques beyond PTSD treatment refer to medical practices conducted without a PTSD diagnosis, i.e., medical practices not covered by public insurance. Therefore, this represents the use of medical techniques for the purpose of enhancement. There may certainly be advantages to manipulating memories for those who do not suffer from PTSD. Imagine the case of Person D and Person E. In the past, D had a heated argument with friend E. Even now, D dislikes E. D has difficulty behaving in a carefree way when sharing space and communicating with E, and thus behaves in a way that lacks initiative. No physician would diagnose D with PTSD. Yet, if possible, D would prefer to manipulate the memories of the past argument in order to suppress the emotional reaction toward E. By doing this, D desires to behave in a more carefree and active manner when communicating with E. D's desire does seem to be understandable.

In addition to closely examining memory manipulation technology from the standpoint of safety, it must also be subjected to ethical scrutiny from the standpoint of personal identity. As mentioned before, however, we focus only on the social value of memory here. The question we must ask is: Is it socially permissible to manipulate episodic memory with social value for the purpose of enhancement even if such manipulation undermines the value or even eliminates it? We believe it is, because an individual has ownership over his or her own memories. Ownership in this context is a concept similar to intellectual ownership like an idea created by a person. Regardless of

whether or not a given episodic memory possesses social value, the will and freedom of the individual based upon such intellectual ownership who undergoes memory manipulation must be respected. The permissibility of altering episodic memory should not be determined based on whether there is a diagnosis, or whether the purpose is enhancement, but should instead be left entirely to the individual's needs.

Yet, there are also episodic memories with social value for which manipulation is impermissible. Imagine another person, F, who is a murderer. For F to manipulate or erase the memory of the murder would unjustly advantage F when giving testimony at trial. At the very least, erasing the memory would provide some sort of advantage to F. Manipulation of this type of memory cannot be socially permitted. It is not simply that F is immoral for committing the crime, but also that the act of attempting to erase that memory is even more immoral since it is kind of destruction of evidence. Thus, when the episodic memory with social value is related to an illegal or unethical behavior, the manipulation of such memories must not be permitted.

2.3. Social Value of Personal Memory and Collective Memory

An episodic memory possessed by an individual is unique to that individual, and the individual alone has privileged access to the memory's contents. Thus, when an episodic memory possessed by an individual has social value, its social value is endorsed by sharing the contents of that memory with others through language. In this manner, such memories are conceptualized and become collective memory. As discussed previously, through compassion, the non-cognitive contents of memory can also modify collective memory. We have argued that in the case of personal memory, memory manipulation should be left up to the free will of the individual. When it comes to collective memory, however, memory that holds social value must not be manipulated arbitrarily. Imagine a new technology was developed to manipulate collective memory on a large such as a past war. We can find its precedents in neurological technologies that are already partially implementable through media that make use of subliminal effects, as well as the older method of 'school education.' However, from the standpoint of protecting the property of humanity, it would be immoral to use such manipulation to obliterate the memory of a past war and other large collective memories.

2.4. Virtue of Preserving Memories with Social Value

If an individual conscious of the social value of his or her memory opts not to go through with manipulating it, even though preservation of the memory brings up negative emotions, then he or she is praiseworthy. In other words, his or her choice is supererogatory. Imagine the case of another person, Person G. G is a firefighter who in the past encountered many tragic fires. G experiences feelings of sadness every time he or she recalls the memory of past fires. Yet, recalling these memories is not accompanied by pathological symptoms. G believes that maintaining as much detail as possible of these memories is valuable for literacy efforts that aim to increase the public awareness of fire safety measures. G has the right to freely control this episodic memory. By choosing not to alter these memories, however, G contributes to society as a narrator. G's actions deserve praise in moral terms. Acknowledging the memory has social value and preserving it, and then narrating it, is supererogatory.

3. Conclusion

Episodic memory can possess social value, and memory manipulation technology has the potential to violate this value. The use of memory manipulation technology to treat PTSD is legitimate. Moreover, even if the purpose is for enhancement, memory manipulation technology should be respected as an exercise of individual freedom, and therefore it follows that the disruption of memory's social value should be permitted. Immoral memories, however, are an exception, and the revision of collective memory requires much caution. In general, the preservation of memory is supererogatory, and the act of those who preserve episodic memory with social value is worthy of praise. Recognizing the social nature of memory, and endorsing, preserving, and narrating it, is supererogatory.

References

Alea, Nicole, and Susan Bluck. 2003. "Why are you telling me that? A conceptual model of the social function of autobiographical memory." *Memory* 11 (2): 165–78. doi:10.1080/741938207

Astill Wright, Laurence, Marit Sijbrandij, Rob Sinnerton, Catrin Lewis, Neil P. Roberts, and Jonathan I. Bisson. 2019. "Pharmacological prevention and early treatment of post-traumatic stress disorder and acute stress disorder: a systematic review and meta-analysis." *Translational Psychiatry* 9 (1): 334. doi:10.1038/s41398-019-0673-5

Bisson, Jonathan I, Lucy Berliner, Marylene Cloitre, David Forbes, Tine K. Jensen, Catrin Lewis, Candice M. Monson, Miranda Olff, Stephen Pilling, David S. Riggs, Neil P. Roberts, and Francine Shapiro. 2019. "The International Society for Traumatic Stress Studies new guidelines for the prevention and treatment of posttraumatic stress disorder: Methodology and development process." *Journal of Traumatic Stress* 32 (4): 475–483. doi:10.1002/jts.22421

Ishikawa, Rie, Chiaki Uchida, Shiho Kitaoka, Tomoyuki Furuyashiki, and Satoshi Kida. 2019. "Improvement of PTSD-like behavior by the forgetting effect of hippocampal neurogenesis enhancer memantine in a social defeat stress paradigm." *Molecular Brain* 12 (1): 68. doi:10.1186/s13041-019-0488-6

Koizumi, Ai, Kaoru Amano, Aurelio Cortese, Kazuhisa Shibata, Wako Yoshida, Ben Seymour, Mitsuo Kawato, and Hakwan Lau. 2016. "Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure." *Nature Human Behaviour* 1:0006. doi:10.1038/s41562-016-0006

Kolber, Adam J. 2006. "Therapeutic Forgetting: The Legal and Ethical Implications of Memory Dampening." *Vanderbilt Law Review* 59 (5): 1561–1626.

Lacy, Joyce W., Craig E. L. Stark. 2013. "The neuroscience of memory: implications for the courtroom." *Nature reviews. Neuroscience* 14 (9): 649–658. doi:10.1038/nrn3563

Lavazza, Andrea. 2018. "Memory-modulation: Self-improvement or self-depletion?" *Frontiers in psychology* 9: 469. doi: 10.3389/fpsyg.2018.00469

Lavazza, Andrea. 2019. "Moral bioenhancement through memory-editing: A risk for identity and authenticity?" *Topoi* 38: 15–27.

Liao, S. Matthew, and Anders Sandberg. 2008. "The normativity of memory modification." *Neuroethics* 1, 85–99. doi:10.1007/s12152-008-9009-5

Steingrimsson, Steinn, Gorana Bilonic, Ann-Catrin Ekelund, Tomas Larson, Ida Stadig, Mikael Svensson, Iris Sarajlic Vukovic, Constanze Wartenberg, Olof Wrede, and Susanne Bernhardsson. 2020. "Electroencephalography-based neurofeedback as treatment for post-traumatic stress disorder: A systematic review and meta-analysis." *European Psychiatry* 63 (1): e7. doi:10.1192/j.eurpsy.2019.7

Tulving, Endel. 1972. "Episodic and semantic memory." In *Organization of memory*, ed. E. Tulving and W. Donaldson, 381–403. New York: Academic Press.

van der Kolk, Bessel A, Hilary Hodgdon, Mark Gapen, Regina Musicaro, Michael K. Suvak, Ed Hamlin, and Joseph Spinazzola. 2016. "A Randomized Controlled Study of Neurofeedback for Chronic PTSD." *PLoS One* 11 (12): e0166752. doi:10.1371/journal.pone.0166752

Journal of Cognition and Neuroethics

The Mental & Physical Health Argument Against Hate Speech

John Park
California State University, Sacramento

Biography

John J. Park is an assistant professor of philosophy at California State University, Sacramento. He received his PhD from Duke University in philosophy. His research interests are in ethics, political philosophy, and philosophy of mind & cognitive science. He has published the book *The Psychological Basis of Moral Judgments* with Routledge.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2023. Volume 9, Issue 1.

Citation

Park, John. 2023. "The Mental & Physical Health Argument Against Hate Speech." *Journal of Cognition and Neuroethics* 9 (1): 13–34.

The Mental & Physical Health Argument Against Hate Speech

John Park

Abstract
Overall, there's a rich literature on free speech and hate speech. However, there's been comparatively less discussion on hate speech that brings in empirical psychological and medical evidence on the possible health harms hate speech can have for minorities. I introduce and piece together a set of pre-existing scientific data that's new to the philosophical literature to help sufficiently establish an argument that governments should ban hate speech. Given the adverse effects hate speech can have on one's mental and physical health, hate speech causes harm at many times. Since it causes such harm and leads to overridingly negative consequences, the government ought to regulate such speech.

Keywords
Empirical Moral Psychology, Hate Speech, Free Speech, Political Philosophy

1. Introduction

The issue of free speech occupies an important place in political philosophy. The importance of this issue is even more pressing given the somewhat recent rise of the alt right movement in the U.S. and increase of hate speech on the internet. The question is not whether speech should be limited by the government. Philosophers generally agree that speech should be limited for instances like inciting violence, libel & slander, lying that there's a bomb on a plane, false advertising, blackmail, advertising dangerous adult products to children, child pornography, uttering "fighting" words, lying by yelling 'fire' in a crowded theater, etc. Free speech is not a basic infallible value. The real question is how much should speech be limited. A particular controversial issue that is of singular interest here is whether hate speech should be limited by the government. We can define the term 'hate speech' generally as public communication expressing abusive prejudice and/or violence towards a person or group based on race, religion, gender, ethnicity, disability, age, or sexual orientation.

There are divergences in political regulations on hate speech even within Western liberal democracies. For example, the United States is on the upper end of the spectrum of putting more weight in favor of free speech rather than limiting hate speech.

Meanwhile, places like Germany, the U.K., Poland, France, Hungary, and Austria have hate speech laws in significant part due to the history of Nazis, the Holocaust, and World War II on European soil. Others like Canada, Australia, and Mexico also have hate speech laws. Indeed, regarding hate speech, the U.S. is more of the exception in allowing for it rather than the rule.

Casting a long shadow over the hate speech debate is Mill's (1859/1978) classical work on liberty and the harm principle. The harm principle in light of hate speech says that we are allowed to generally speak freely but harming others provides a pro tanto reason for regulating speech. Mill appears to maintain that harming others is not a sufficient reason for regulating speech. In order to have a sufficient reason, further justification is needed for limiting speech. Mill, who is a consequentialist, says that one must show that the consequences of the harm outweighs the consequences of the benefits of free speech in the given instance. It's this generally accepted or orthodox approach in ethics of one means for when speech and liberty should be limited that's grounded in the harm principle that will be the main mechanism used for our inquiry as to whether hate speech should be regulated by the government.

Moreover, Mill states that the harm principle can be applied prospectively to help prevent future harm. The risk of harm allows for the invocation of the harm principle even if the harm may not come about. For example, he writes in *On Liberty* that we may limit the liberty of individuals from being intoxicated at work since it elevates the risk of workplace accidents and physical harm to fellow employees. In modern times, the government may limit the ability to drive while intoxicated given the risk of bodily injury to others even though no accident may occur in a particular instance. Such limitations are justified even though no harm actually does ensue on a case.

What is a 'harm' is a controversial issue debated by philosophers. While this issue would constitute its own separate paper, we may continue our inquiry for our purposes by relying on a current highly influential understanding of 'harm' given by Joel Feinberg (1984). 'Harm' is not meant to refer to any kind of damage to any object whatsoever, but it is about a wrongful setback to the interests of a person.[1] 'Interests' are one's stake

1. I focus on Feinberg's account of harm as his is an influential view in contemporary philosophy. Mill differs in that he defines 'harm' as a setback of one's important interests *in which one has rights*. 'Rights' for Mill ultimately are grounded in utility and promote good consequences. Mill believes that we have a right to be free from unwarranted bodily injury from others. For example, he states that if the buying of poison had no other purpose than to murder others, then it should be banned. He says that drinking on the job should be banned because one might hurt others at work. Insofar as I show that hate speech can cause physical injury to others, I also demonstrate that hate speech is a harm on Mill's account.

on an issue, where one's life may go better or worse in the long run depending on what happens. Feinberg presents examples of interests like career development, keeping one's property, maintaining the well-being of one's family, having social justice for one's community, and having one's country be economically secure. An example of what is not a harm is when one's favorite sports team loses a game. This isn't a setback of one's interests as it is merely a temporary discomfort or pain. If something negative will go away and leave an individual as they were previously, whole and undamaged, then one's interests were not set back. Yet, perhaps there might be fanatics who do become mentally ill in the long term because their team didn't win the championship. Even so, the opposing team beating one's team isn't in-itself morally wrongful. It's just fair athletic competition. Therefore, it isn't a harm. Here, we see the importance of 'wrongful' in an analysis of 'harm.'

Feinberg writes:

> [A]n affront or an insult normally causes a momentary sting; we wince, suffer a pang or two, then get on with our work, unharmed and whole. But if the experience is severe, prolonged, or constantly repeated, the mental suffering it causes may become obsessive and incapacitating, and therefore harmful. (1984, 45–46)

Feinberg does not classify hate speech as a harm but as a mere offense as he believes it has just a temporary sting on minorities without any lasting mental or physical damage (1985). Minorities are still left whole in the long term after being insulted with racial slurs, such as when black people are called the N-word. However, notice in Feinberg's quote that if certain wrongful speech is found to have long term effects on mental health, then such speech is a harm. The harm principle comes into play. It will be one goal among others in this paper to demonstrate that wrongful hate speech does at many times cause damage to one's mental and physical health. Therefore, once I demonstrate this, we can state unequivocally that hate speech is a harm.

Mill justifies liberty in speech for many reasons. Some of them are that we should have a marketplace of ideas because a silenced opinion can be true, a silenced opinion importantly may contain a partial truth, and speech that challenges what is true allows us to test the truth so that the truth is not held in mere prejudice that's inherited rather than adopted. However, Mill allows for restrictions on freedom of speech in part with the harm principle. Today, Mill's harm principle-based route is perhaps the predominant means in philosophy for the government potentially limiting liberty and certain speech, although there may be alternative means for doing so as well.

There are numerous contentions against hate speech (Langton 1990, Fish 1994, Waldron 2012, McGowan 2012, Maitra 2012, Brown 2015). For example, Laura Beth Nielsen (2012) draws on empirical studies to show that the proper response to hate speech is not to allow it and advocate for more speech that counters hate speech. For, data demonstrates that minorities subject to hate speech overwhelmingly desire to keep silent after being subjected to hate speech due to numerous factors such as fear from physical harm. Caroline West (2012) contends that hate speech should be banned because empirical data shows that it serves to generally undermine and silence speech rather than enhance it.

Others like David Boonin (2011) argue against regulating hate speech. For example, he says that at many times hate speech can be viewed as fighting words. Since we already have regulations on fighting words, we don't need new regulations on hate speech. Moreover, as we shall continue to see shortly, it's the received view in philosophy that hate speech isn't a health harm and doesn't cause long term damage to health. The harm principle can't be invoked for this reason. Therefore, it's questionable as to whether there needs to be government involvement on it.

While there may be many good arguments above against hate speech, our purpose here is not to assess or expand upon any of such previous contentions against hate speech. Instead, it is to solely focus on a different avenue concerning the effects hate speech has on mental and physical health. Through this route, I contend that hate speech ought to be banned.

2. Criticisms of Other Attempts To Use Health Harms

There have been a few scholars who have discussed the mental effects and health harms hate speech may have in order to argue against hate speech. However, these only have been very small in number, and they only have provided speculative or insufficient cases. I now offer my own criticisms of them. Richard Delgado (1993) speculates that hate speech along with institutionalized discrimination can cause adverse immediate mental effects, such as fear, stress, and anger. However, they might actually have long term effects. They may impact health in that they may lead to mental illness, such as depression, or physical illness, such as high blood pressure. Nevertheless, Delgado's essay lacks rigorous empirical data to establish the health effects hate speech might have, especially given that empirical psychological work on hate speech and discrimination in the early 90's was still a budding field.

Melina Bell (2021) makes a recent attempt to show that hate speech can cause psychological harm in order to contend that hate speech should be banned. However, it is insufficient. She cites studies where women expressed a greater state of self-objectification after viewing comedy skits that objectified women as sex objects (Ford et al. 2015). She then states that self-objectification can lead to mental illness. However, one concern is that these were not longitudinal studies. While the jokes may have caused self-objectification in the short term, it may have been short-lived because the women may have at least subconsciously thought later that it's really only a joke. Professional comedic settings where statements usually are made in jest may provide a different circumstance in the minds of women in the long run. Many women may not stay worked up about it over time, and many may develop mental illnesses over gender discrimination from non-comedic contexts rather than from comedic ones. This is a reasonable possibility, so replicated longitudinal experiments are needed to show that self-objectification from comedy skits contributes to long term health issues rather than only hate speech that's given in non-comedic contexts. Without such studies, Bell's contention is invalid.

Moreover, it also could be the case that relevant speech in comedy skits may cause self-objectification at least initially since others are laughing and agreeing. Meanwhile, hate speech in non-comedy settings might cause no illnesses. One can't necessarily infer from these studies in comedic contexts that hate speech in non-comedy settings causes mental illnesses in many women. Given the context of comedic settings, studies on the effects of discriminatory speech in non-comedy settings are needed in order to sufficiently and decisively show that hate speech in non-comedy settings causes harm.

Bell also relies on studies suggesting that discrimination leads to health problems like heart disease, cancer, and strokes for African Americans (Lewis & Van Dyke 2018). The studies importantly were able to weed out factors like socioeconomic status, health behaviors, and access to care as factors for causing such problems. This makes it more likely that discrimination is the cause of such maladies. However, the studies don't eliminate other important possible confounding variables that may be responsible for the illnesses rather than discrimination. For example, birthplace and education level were not eliminated as possible causal influences for the negative health outcomes. More data is needed to eliminate such possible confounding variables in order to have a valid argument that discrimination and hate speech cause such health illnesses. Overall, Bell is unable to sufficiently show that hate speech is a harm and causes health illnesses. To note, in my own argument below, I will present studies that eliminate the above confounding variables.

In a *Stanford Encyclopedia of Philosophy* entry on freedom of speech, David van Mill notes that regarding hate speech and a Neo-Nazi march in Skokie, Illinois:

> [W]e might want to claim that [the minorities] were psychologically harmed by the march. This is much more difficult to demonstrate than harm to a person's legal rights. It seems, therefore, that Mill's [harm principle] does not allow for state intervention in this case (2017).

There's a general sense in philosophy that the empirical evidence for hate speech causing mental illnesses is not established as of yet, or at the very least, it's controversial.

Works in philosophy advocating for an end to hate speech by taking the avenue of medical health have yet to put forth a systematic and sufficient case that hate speech has significant negative effects on mental or physical health. Thus, I will attempt to be the first in the philosophical literature to rely on modern empirical findings in psychology and medicine to sufficiently establish such negative outcomes and then argue for legislative measures against hate speech. One of my contributions is to introduce to the philosophical literature on hate speech the sufficient empirical evidence of how hate speech can cause negative health outcomes. I then argue that hate speech generally should be banned by the government in part given the effects it can have on human health. This is a general moral rather than legal claim I'm making here that the government ought to regulate hate speech, and I leave the later question of what specific policy measures should be implemented by the government to ban hate speech for a later time. I only make a general moral claim here that hate speech ought to be banned.

3. Empirical Psychology Evidence on Hate Speech & Health

The first step of positing my contention against hate speech is to establish that hate speech is a statistically significant factor for causing negative mental and physical health consequences for many minorities. This is a causal empirical claim that hate speech causes health harms. It's not a constitution claim that hate speech itself is a harm, such as when a president utters a discriminatory mandate that itself enacts policy that is harmful. In this instance, the speech constitutes harm since it enacts a harmful policy. While I don't deny the possibility that hate speech also can constitute harm, the focus of my argument is only on a specific causal claim that hate speech can cause harm; namely, to one's health. This causal rather than constitution claim will be what I need to establish my thesis.

Social scientists and researchers in the medical field have amassed hundreds of studies demonstrating that hate speech and other kinds of discrimination can have adverse effects on mental and physical health. Discrimination is a broader category of which hate speech is a token. To note, I rely on studies that specifically involve hate speech along with experiments that involve many other kinds of discrimination. The results are similar for both. However, as hate speech is a type of discrimination, replicated negative health outcomes found for diverse kinds of discrimination provide inductive support that they will apply to hate speech as well. Inductive logic is the logic of probabilities, where a good inductive argument with true premises will support the conclusion with a sufficient degree of likelihood rather than by 100% certainty. If many different kinds of discrimination cause bad health outcomes, then this increases the likelihood that hate speech is causally efficacious too in leading to such outcomes. This is like how discovering that many diverse species of primates practice some level of cooperation provides good reason to believe that a disparate newly discovered primate species also practices cooperation. As many different mRNA-based vaccines in history have proven to be safe in the long term, this means that the new COVID-19 mRNA vaccines, such as by Pfizer and Moderna, will likely be safe in the long term despite an absence of longitudinal studies on the new vaccines. This basic use of inductive logic for my conclusion is further strengthened based on the fact that studies to be discussed below that directly test for the effects of hate speech also are shown to lead to similar health outcomes. Insofar as replicated evidence on hate speech is also utilized, this staves off the objection that hate speech doesn't cause health maladies because experiments just on discrimination that don't involve hate speech can't be used to make such an inference. Moreover, the use of inductive logic, as described above, utilizing other diverse kinds of discrimination also sufficiently addresses this objection.

The existence of hate speech along with data showing that institutionalized discrimination for practices like job hiring (Quillian et al. 2017), housing (Schneider 2018), and criminal justice (Sentas 2018) are still going strong provides an overall environment that can be deleterious to the health of many minorities. It can lead to chronic stress that causes serious health harms.

First, there is psychological evidence that hate speech and discrimination correlate with stress (Krieger & Sydney 1996, Ren et al. 1999, Finch et al. 2000, Gelber & McNamara 2016). Many of these studies are interviews where hate speech and discrimination seemingly co-occur with stress. However, they as of yet don't establish causation that hate speech causes stress. One can't establish causation from mere personal phenomenological reports and psychological interviews since hidden subconscious mental

events that really drive mental processes may be different from conscious awareness and recollection. Rather, more rigorous experimental methods are required to help establish causation. Regardless, such correlational studies are important and potentially still help to contribute to a causal claim because you can't have causation between two events unless you have correlation.

As an example of a correlational national study, African Americans, Hispanic Americans, and Asian Americans reported higher levels of stress than whites supposedly in significant part due to experiencing racism (Williams 2000). In another study, 96% of participants who experienced racism within a past year reported feeling stressed from the incident (Klonoff and Landrine 1999). Nielson (2012) conducted an ethnographic study interviewing those who experienced hate speech. She found that many reported experiencing stress and fear for their safety from such incidents. For instance, a Filipino woman was interviewed and recalled an episode.

> And I was at a gas station, and a guy came out, didn't talk to me directly, but I knew he was talking about me. I was seated in the car, and the driver who was beside me was white, and the guy just said, kind of in the air, "I can see the driver's the only human being around here." Implying I was not a human being.
>
> Q: Uh huh. And did you respond to that in any way?
>
> A: No, I didn't because I was afraid. (2012, 156)

In another case, a young woman was interviewed.

> I know just last week, I was in the BART station at Montgomery and there was, um, I think a homeless man who came up to me and said, "I hate women, they're all sluts"…That probably sticks in my mind the most…
>
> Q: Um, what did you say to the guy who, um, informed you that all women are sluts?
>
> A: Um, I just turned around; I didn't say anything. I was pretty scared of him. (2012, 161)

A different study based on observation from a psychologist found an association between psychological trauma with an instance of racial discrimination.

> A light-skinned Hispanic male was treated courteously when he made application for an apartment in New York City. However, when he returned with his African-American wife, the renting agent became aloof and informed them that the apartment was rented. In response to the denial of the apartment, the wife immediately became depressed, insomniac, and hypervigilant. She had repeated nightmares. At the time of the alleged discrimination, she noticed that her hair had begun to fall out, that her skin was dry, and she was constipated…there was a mild paranoid trend. All of her symptoms were causally related to the discrimination. (Butts 2002, 338)

In a meta-analysis of 333 published articles which involves a total sample across all studies of 309,687 participants, experiencing racism was associated across a variety of minority groups with mental health issues, such as depression, anxiety, posttraumatic stress disorder (PTSD), and suicidal ideation (Paradies et al. 2015). The results were statistically significant. Furthermore, it was significantly correlated with poorer physical health.[2] They write that "[t]his meta-analysis indicates that racism is significantly related to poorer health…" (Paradies et al. 2015, 24).

What's interesting about this meta-analysis of studies including others (Lewis & Van Dyke 2018) is that they also provide support for a causal claim that hate speech can cause mental and physical health ailments given that many other possible causal factors for the health ailments were eliminated. Through statistical analysis, the researchers found that factors like age, sex, birthplace, socioeconomic status, health behaviors, and education level didn't moderate the effects of racism on health. It looks like experienced racism rather than other factors is what's really causing health problems. Paradies et al. (2015, 28) draw a causal conclusion that "racism has long-term effects on health that remain significant despite attenuation over time." They also point out that given our knowledge of childhood psychology, exposure to racism like hate speech early in life can make one more vulnerable to its health effects with more severe and persistent health consequences given the biological embedding of early life stress.

Moreover, there's a well-established and uncontroversial literature in psychology and medicine on Adverse Childhood Experiences (ACE). By also eliminating other possible confounding factors, they decisively show that certain negative traumatic childhood experiences like witnessing violence or facing abuse from 0 to 17 years of age can cause

2. This correlation fell under the category of miscellaneous physical health.

early mortality, chronic health problems like heart and liver disease, and mental illnesses in victims (Anda et al. 2002, Dong et al. 2004, Brown et al 2010). Part of what's included in ACE is hate speech and racist experiences (Pachter et al. 2009, Wade 2014, Bernard 2021). The ACE studies in themselves establish a strong causal connection of hate speech causing adverse mental and physical health outcomes in youths.

As another study that helps to establish causation, perceived racism by African American women throughout their lifetime accurately predicts their infants' birthweight beyond the effects of medical and sociodemographic factors (Dominguez et al. 2008). Low birthweight babies commonly have complications in fighting off infections, breathing, staying warm, and feeding to gain weight. They are at elevated risk for cerebral palsy, blindness, and deafness. Given that experienced racism is such an accurate predictor of birthweight and many other potential causal factors for decreased birthweight have been eliminated, this suggests that racist acts like hate speech have a deleterious effect on the health of relevant infants. Moreover, suicide rates for ethnic immigrant groups in the U.S. are significantly predicted by the degree of negativity of hate speech they experience (Mullen & Smyth 2004).

Hate speech also can lead to negative physiological changes (Krieger 1990, Krieger and Sidney 1996, Clark et al. 1999, Guyll et al. 2001, Brondolo et al. 2003, Williams et al. 2003, Harrell et al. 2003, Bennett et al. 2004). Physiological changes include activation of the hypothalamic pituitary adrenal cortical system, which then leads to the release of cortisol, a stress hormone. It also entails alterations in immune and cardiovascular functioning. Cardiovascular functioning as well as respiratory functions and pupil size come under control of the sympathetic branch of the autonomic nervous system during emergencies. These physiological changes make up the sympathetic adrenal medullary axis of the stress response, where laboratory and survey evidence link such responses to subjects who experience discrimination (Brondolo et al. 2003, Harrell et al. 2003).

For example, in psychophysiological investigations, subjects experience racially charged encounters in laboratory settings, where physiological effects are measured (Morris-Prather et al. 1996, Kinzie et al. 1998, Blascovich et al. 2001). The racially charged encounters include situations of experiencing hate speech among other discriminatory acts. These are compared to non-racist encounters as controls. Such experiments permit the drawing of cause and effect conclusions about the relationship between hate speech and physiological changes. From overviewing psychophysiological experiments, Harrell et al. state "that direct encounters with discriminatory events contribute to negative health outcomes (2003, 243)."

A moderated psychophysiological approach uses personality measures as predictors of disparities in physiological responses between subjects to racist events. If relevant personality trait differences reliably lead to divergences in physiological responses, then this provides causal evidence that hate speech causes negative physiological harm in relevant participants. For instance, experiments show that those who are more passive in addressing the discrimination they experience have higher blood pressure than those who are more active (Krieger and Sidney 1996). African American women with passive coping responses were 4.4 times more likely to have hypertension or high blood pressure than African American women with active strategies (Krieger 1990). Moreover, diastolic blood pressure reactivity to a speech stressor was augmented for participants who had previously experienced discrimination as compared to those who didn't (Guyll et al. 2001). Hypertension itself is a serious condition that is a leading cause of heart attacks and strokes.

As an alternate argument, the general stress literature can be utilized. Given that hate speech causes stress for minorities, being subject to such speech along with institutionalized discrimination has serious consequences for one's health. At this point in the psychological literature, it's uncontroversial that hate speech can be a very stressful event, where a majority of those who experience hate speech feel stress (Kessler et al. 1997, Clark et al. 1999). It's uncontroversial that hate speech is a significant cause of stress, and it's present within societies with institutionalized discrimination. This environment leads to chronic stress for many individuals. Once we have this on the table, we then can rely on the general stress literature to show that this likely will cause many to suffer poorer health. The harmful effects chronic stress can have on mental and physical health is well documented and uncontroversial. The National Institute of Mental Health states:

> [Chronic stress] can disturb the immune, digestive, cardiovascular, sleep, and reproductive systems…Over time, continued strain on your body from stress may contribute to serious health problems, such as heart disease, high blood pressure, diabetes, and other illnesses, including mental disorders such as depression and anxiety. (2020)

This somewhat indirect route that relies on the general stress literature also can be used to establish and buttress the causal claim of the effects hate speech can have on health, although we have discussed above more direct studies showing that hate speech causes many to suffer some of these serious health maladies.

The above kinds of studies help to show the negative influence hate speech has on health. Given, the many replicated studies that eliminate other possible causal variables that can negatively impact health, ACE studies, psychophysiological investigations, moderated psychophysiological experiments, the stress literature, and correlational findings, I believe it's safe to conclude that hate speech is a statistically significant factor for leading to harmful mental and physical health outcomes.

To be sure, whether one experiences damage to one's health will depend on the individual. For example, persons with different ways to cope with hate speech can have different mental and/or physical outcomes. Factors like genetics, personality, childhood experiences such as experiencing hate speech early on, and support networks such as having family or being part of a religious group, can affect what health outcomes one will have through experiencing institutionalized discrimination and hate speech throughout one's life. However, the data still shows that hate speech and other acts of discrimination are statistically significant causal influences for certain negative health outcomes. It's important to keep in mind that even though damage to health can be contingent upon the experiencer of hate speech, those who commit the discriminatory act still do damage to another's health and are causally responsible. This is somewhat similar to how an assaulter hitting another is causally responsible for the ensuing physical harm even though the assailed wasn't dexterous enough as some others to dodge the attack.

I take it that there's a general consensus on the effects of hate speech in the scientific literature. The Surgeon General reports that "the findings indicate that racism and discrimination are clearly stressful events...Racism and discrimination adversely affect health and mental health" (U.S. Department of Health and Human Services 2001, 38) of minorities and places them at risk for mental disorders. In law, given the scientific data, claims of psychological distress in reaction to hate speech have been viewed as credible injuries (Neu 2008).

4. The Mental & Physical Health Argument Against Hate Speech

Now that we've established that hate speech is a statistically significant cause of adverse health effects, I will contend that this helps to show that hate speech ought to be banned by the government. Recall that harms are wrongful setbacks to one's interests, and this includes long term damage to one's physical and mental health. We have seen how hate speech is a setback in that it can lead to long term damage of one's mental and physical health in non-comedic settings. Furthermore, uttering hate speech against others

is in-itself wrong to do as it expresses abusive discriminatory insults or violence to certain groups. It is a kind of unjust discrimination disparaging or threatening people based on morally arbitrary factors like the color of one's skin. Somewhat similar to how a child is wrong for name-calling on the playground, hate speech is a wrong. Given the above, we can utilize the generally accepted route that's based on the harm principle. We can say as a first pass that speech that causes harm runs into the harm principle, and this provide a pro tanto reason to ban such speech.

Given my new establishment in the philosophical literature that hate speech can cause health injuries, it could be that many kinds of speech may be stressful to others and may cause health harms, such as possibly telling your children to study hard and do well in school. However, telling your children to study hard isn't in-itself wrong to do. Thus, it isn't a harm. Also, recall that when invoking the harm principle, we need to do a consequentialist weighting of the harms and benefits in order to potentially be sufficiently justified in regulating speech. Let us call (1) the application of the harm principle and (2) this consequentialist analysis that lies in favor of regulating action *the harm principle conditions*. Satisfaction of the harm principle conditions is sufficient for banning certain speech. With this in mind, the greater overall good for society is for people to tell their children to study to produce more productive future citizens, so the second condition also isn't satisfied in the case of telling your children to study. Given the above, instructing your children to study shouldn't be banned by the government.

As I already have established the first component that hate speech invokes the harm principle, let us now assess the second component of the harm principle conditions. We need to examine whether the overall consequences favor the benefits of being able to utter hate speech or whether they give more weight to the harms of minorities suffering from death, mental illness, and physical ailments caused by hate speech.

First, the benefits of hate speech are very low. Discriminatory policies can be presented in a civil manner rather than hurling insults and/or physical threats at others. Mill's benefits of free speech concerning uncovering the truth, a partial truth, and making sure we don't hold beliefs dogmatically can be upheld by discriminatory people as they can present their ideas in a civil tone. Uttering hate speech is supererogatory to being free to get across ideas. Hence, Mill's benefits of free speech cannot be used to support hate speech in a consequentialist calculation. In a court of law for a trial, both sides can get across their ideas and arguments without throwing personal insults and/or physical threats to violence at the other party, where judges *will* regulate such abusive vituperations. Likewise, discriminatory people can present their ideas without uttering abusive or threatening hate speech.

We have seen how the benefits of hate speech are minimal. On the other hand, the harms of many people suffering from death, mental illnesses, and physical ailments due to hate speech are tremendous. We know that chronic stress at many times leads to negative physical health outcomes. Recall that hate speech is a statistically significant cause of a variety of physical health illnesses, like having unhealthy infants whose lives are threatened and hypertension which is a leading cause of heart attacks and strokes. Suffering from mental illnesses like depression, PTSD, and anxiety is painful. Mental illnesses can have negative effects on mobility, on the ability to perform physical tasks or even get up out of bed, on one's job performance that impacts subsistence for one's family, on family and friends relationships, on personal well-being, and on pursuit of one's goals in life. Mental health illnesses also can lead to death such as by way of suicide.

Given the little benefits of being able to utter hate speech and the strong negative effects on health of many of those who experience hate speech, the greater consequences are to ban hate speech. We now have seen how hate speech violates the harm principle and how the greater consequences are to ban hate speech. Both components of the harm principle conditions have been satisfied. Therefore, I conclude that the government ought to regulate hate speech. The main contribution to the literature of my essay is that I'm the first to sufficiently argue for the general banning of hate speech by 1) introducing into the philosophical literature a wealth of modern empirical findings that together sufficiently establish that hate speech is harmful to one's health and 2) contending that such health harms in 1) outweigh the little benefits of being able to utter hate speech.

One might object that if we ban hate speech, this will lead to a slippery slope where we'll ban many other different kinds of speech. Soon we'll be living in an authoritarian type of state that has very limited speech such that we can't even criticize the government, and there won't even be freedom of the press. However, unless one establishes that it's causally likely that we'll slip down this slope to an authoritarian-like state concerning speech, one has committed the slippery slope logical fallacy. The burden of proof is on the objector to establish such a causal connection.

Nevertheless, let me point out that there's a stopgap in that my contention against hate speech falls within the harm principle conditions. The harm principle conditions are the stopgap here. I agree that speech that satisfies Mill's harm principle conditions should be outlawed by the government, and this is really as far as I need for my thesis. Therefore, this still perfectly allows for speech that criticizes the government, and it allows for freedom of the press since the harm principle conditions generally don't apply to such speech. Given this stopgap of the harm principle conditions, there's no further reason to think we'll slip down the slope.

Another possible counter is that there is a less costly alternative to government regulation on hate speech that can eliminate the health harms, so there shouldn't be the relevant ban. We should support victims of hate speech to give more speech that counters the hate speech they experienced. This might act as a stress relief such that they don't experience negative health effects. However, as already mentioned, data shows that minorities are not likely to participate in such counter-speech due to factors like fear and intimidation (Nielson 2012, West 2012). After all, historically, there is a credible link between certain instances of hate speech and acts of physical violence against minorities.

David Boonin (2011) argues that there's no need for legislation on hate speech at all. Hate speech is not immediately threatening towards physical violence and isn't immediately "fighting" words because if an elderly and frail white grandmother in a wheelchair said such words, they wouldn't be physically threatening nor stir a minority to fight. Thus, in this respect, there can't be a blanket ban on hate speech that, for example, restricts the grandmother. Also, in contexts where hate speech is threatening towards physical violence or is a "fighting" word, then we already have regulations on such speech, so no new laws are needed. Ultimately, there should be no laws put forth restricting hate speech. However, I have shown that hate speech causes harm to one's mental and physiological health and leads to an overall disutility. Even such words uttered by a frail grandmother can have this effect. Since the harm principle conditions are satisfied, there should be a general legislative ban on hate speech that even covers cases like the frail grandmother.

As another possible objection, one may say that my argument might then allow for banning non-hate speech that causes damage to one's mental or physical health. What about other kinds of unethical speech whose effects are identical to that of hate speech? For example, maybe there should be rules regulating making fun of people who are shorter than average. My response to this is that this might be true, and I would be perfectly happy with the result if so. My contention against hate speech is based in part on the empirical data of the health effects hate speech can have. The fact that there's institutionalized discrimination also exacerbates the situation. This all creates a culture and environment for chronic stress, which leads to negative health outcomes. Moreover, such harms outweigh the benefits. It's in part an empirical question of socio-psychological research whether such an environment or something like it also holds for those who are shorter than average such that they live in a culture that supports chronic stress for them. If the harm principle conditions are satisfied for speech insulting people who are shorter than average, then this provides good reason to outlaw such speech given that it satisfies the harm principle conditions.

An objector may claim that certain speech in the press that causes stress for a criminal might then have to be outlawed on my account, but it shouldn't. For example, on my view, news sources may not be able to print or post articles and editorials condemning a convicted serial child rapist and making it known that an agent is such a criminal. For, this may cause chronic stress for the rapist which can lead to negative health consequences. However, remember the harm principle conditions and my handling of the possible stress caused by telling one's children to study hard. There can be several independent reasons for justifying freedom of the press in this instance. However, for our purposes, it's sufficient to state that writing news stories on crimes isn't in-and-of-itself wrongful. This is especially so when satisfying the normative proposition that the press shouldn't divulge the names of the rape victims in order to protect them. There should be rape shield laws. Since writing news stories on crimes isn't wrongful, this doesn't qualify as a harm. Moreover, it will be for the greater good for society to mete out deserved punishment for such individuals with public rebukes, to inform the public of dangerous people, and to make it known unequivocally that such criminal behavior is undesirable and should not happen. The benefits of such speech from the press outweigh the potential mental harm to the criminal. Thus, my account allows for news sources to print or post accurate articles of and editorials on the child rapist.[3]

In conclusion, it is highly controversial in philosophy whether hate speech should be regulated by the government or not. My contention against hate speech is the first in the philosophical literature to use sufficient evidence from psychology and medicine that shows that hate speech causes mental and physical illnesses. I have shown how this sufficient medical evidence is largely absent in the philosophical literature, and there's the general current thought in the philosophical literature that it hasn't been proven yet that hate speech causes psychological or physical harm. However, I have demonstrated that hate speech is a health harm, and using such data, I also have contended that the health harms outweigh the benefits of uttering hate speech. Therefore, given the satisfaction of the harm principle conditions, hate speech should be banned by the government, and I have defended this thesis from numerous counters. The U.S. should join most of the developed world and regulate such speech that has inimical effects on mental and physical health.

3. There are other objections against hate speech regulations in the literature (Post 2012, Heinze 2016, Strossen 2018). However, I take it that they have been adequately rebutted in the literature, so I don't address them here (Delgado & Stefancic 1996, Brown 2015, Brown & Sinclair 2020).

References

Anda, R.F., Whitfield, C.L., Felitti, V.J., Chapman, D., Edwards, V.J., Dube, S.R. and Williamson, D.F. 2002. "Adverse Childhood Experiences, Alcoholic Parents, and Later Risk of Alcoholism and Depression." *Psychiatric Services 53*: 1001–1009.

Bell, Melina. 2021. "John Stuart Mill's Harm Principle and Free Speech: Expanding the notion of harm." *Utilitas* 33: 162–179.

Bennett, G.G., Merritt, M.M., Edwards, C.L., and Sollers, J.J. 2004. "Perceived Racism and Affective Responses to Ambiguous Interpersonal Interactions Among African American Men." *American Behavioral Scientist* 47: 63–76.

Bernard, D.L., Calhoun, C.D., Banks, D.E., Halliday, C.A., Hughes-Halbert, C. and Danielson, C.K. 2021. "Making the "C-ACE" for a Culturally-Informed Adverse Childhood Experiences Framework to Understand the Pervasive Mental Health Impact of Racism on Black Youth." *Journal of Child & Adolescent Trauma* 14: 233–247.

Blascovich, J., Spencer, S.J., Quinn, D. and Steele, C. 2001. "African Americans and high blood pressure: The role of stereotype threat." *Psychological science* 12: 225–229.

Boonin, D. 2011. *Should Race Matter? Unusual Answers to the Usual Questions*. New York: Cambridge University Press.

Brink, David. 2018. "Mill's Moral and Political Philosophy." *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/mill-moral-political/. Last visited 4/30/21.

Brondolo, E., Rieppi, R., Kelly, K., Gerin, W. 2003. Perceived Racism and Blood Pressure: A Review of the literature and conceptual and methodological critique. *Annals of Behavioral Medicine* 25: 55–65.

Brown, Alexander. 2015. *Hate Speech Law*. New York: Routledge.

Brown, A. and Sinclair, A. 2020. *The Politics of Hate Speech Laws*. New York: Routledge.

Brown, D.W., Anda, R.F., Felitti, V.J., Edwards, V.J., Malarcher, A.M., Croft, J.B. and Giles, W.H. 2010. "Adverse Childhood Experiences are Associated with the Risk of Lung Cancer: A prospective cohort study." *BMC Public Health 10*(1): 1–12.

Butts, H. 2002. The Black Mask of Humanity: Racial/ethnic discrimination and post-traumatic stress disorder. *Journal of the American Academy of Psychiatry and the Law* 30: 336–339.

Clark, R., Anderson, N.B., Clark, V.R., Williams, D.R. 1999. "Racism as a Stressor for African Americans: A biopsychosocial Model." *American Psychologist* 54: 805–816.

Delgado, Richard. 1993. "Words that Wound: A tort action for racial insults, epithets, and name calling. In *Words that Wound*. Edited by M. Matsuda, C. Lawrence, R. Delgado, and K. Crenshaw. Boulder, CO: Westview Press: 89–110.

Delgado, R. and Stefancic, J. 1996. "Ten Arguments Against Hate-Speech Regulation: How valid?" *Northern Kentucky Law Review* 23: 475–490.

Dong, M., Giles, W.H., Felitti, V.J., Dube, S.R., Williams, J.E., Chapman, D.P. and Anda, R.F. 2004. "Insights Into Causal Pathways for Ischemic Heart Disease: Adverse childhood experiences study." *Circulation* 110: 1761–1766.

Feinberg, Joel. 1984. *The Moral Limits of the Criminal Law*. Oxford: Oxford University Press.

Feinberg, Joel. 1985. *Harm to Others*. Oxford: Oxford University Press.

Finch, B.K., Kolody, B., and Vega, W.A. 2000. "Perceived Discrimination and Depression Among Mexican Origin Adults in California." *Journal of Health and Social Behavior* 41: 295–313.

Fish, Stanley. 1994. *There's No Such Thing as Free Speech...and it's a good thing too*. New York: Oxford University Press.

Ford, T.E., Woodzicka, J.A., Petit, W.E., Richardson, K., and Lappi, S.K. 2015. "Sexist Humor as a Trigger of State Self-objectification in Women." *Humor* 28: 253–269.

Gelber, Katharine and McNamara, Luke. 2016. "Evidencing the Harms of Hate Speech." *Social Identities* 22: 324–341.

Guyll, M., Matthews, K.A., Bromberger, J.T. 2001. "Discrimination and Unfair Treatment: Relationship to cardiovascular reactivity among African American and European American women." *Health Psychology* 20: 315–325.

Harrell, C., Hall, S., Taliaferro, J. 2003. Physiological Responses to Racism and Discrimination: An assessment of the evidence." *American Journal of Public Health* 93: 243–248.

Heinze, Eric. 2016. *Hate Speech and Democratic Citizenship*. Oxford: Oxford University Press.

Kessler, R.C., Mickelson, K.D., and Zhao, S. 1997. "Patterns and Correlates of Self-help Group Membership in the United States." *Social Policy* 27: 27–46.

Kinzie, J.D., Denney, D., Riley, C., Boehnlein, J., McFarland, B. and Leung, P. 1998. "A cross-cultural study of reactivation of posttraumatic stress disorder symptoms: American and Cambodian psychophysiological response to viewing traumatic video scenes." *The Journal of nervous and mental disease* 186: 670–676.

Klonoff, E., and Landrine, H. 1999. "Cross Validation of the Schedule of Racist Events." *The Journal of Black Psychology* 25: 231–254.

Krieger, N. 1990. "Racial and Gender Discrimination: Risk factors for high blood pressure?" *Social Science Medicine* 12: 1273–1281.

Krieger, N. and Sydney, S. 1996. "Racial Discrimination and Blood Pressure: The CARDIA study of young black and white adults." *American Journal of Public Health* 86: 1370–1378.

Langton, Rae. 1990. "Whose Right? Ronald Dworkin, Women, and Pornographers." In *Philosophy and Public Affairs* 19: 311–359.

Lewis, T.T. and Van Dyke, M.E. 2018. "Discrimination and the Health of African Americans: The potential importance of intersectionalities." *Current Directions in Psychological Science* 27: 176–182.

Maitra, Ishani. 2012. "Subordinating Speech." In *Speech and Harm: Controversies over free speech*. Edited by I. Maitra and M.K. McGowan. Oxford: Oxford University Press, 94–120.

McGowan, Mary Kate. 2012. "On 'Whites Only' Signs and Racist Hate Speech: Verbal Acts of Racial Discrimination. In *Speech and Harm: Controversies over free speech*. Edited by I. Maitra and M.K. McGowan. Oxford: Oxford University Press, 121–147.

Mill, David van. 2017. "Freedom of Speech." *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/freedom-speech/. Last Visited 2/9/2020.

Mill, John Stuart. 1859/2001. *On Liberty*. Ontario, Canada: Batoche Books Limited.

Morris-Prather, C.E., Harrell, J.P., Collins, R., Leonard, K.L., Boss, M. and Lee, J.W. 1996. "Gender differences in mood and cardiovascular responses to socially stressful stimuli." *Ethnicity & Disease* 6: 123–131.

Mullen, B., and Smyth, J. 2004. "Immigrant Suicide Rates as a Function of Ethnophaulisms: Hate speech predicts death. *Psychosomatic Medicine* 66: 343–348.

National Institute of Mental Health. 2020. "5 Things You Should Know About Stress." https://www.nimh.nih.gov/health/publications/stress/index.shtml. Last visited 1/31/2020.

Neu, J. 2008. *Sticks and Stones: The philosophy of insults*. New York: Oxford University Press.

Nielsen, Laura B. 2012. "Power in Public: Reactions, responses, and resistance to offensive public speech. In *Speech and Harm: Controversies over free speech*. Edited by I. Maitra and M.K. McGowan, 148–173. Oxford: Oxford University Press.

Pachter, L.M. and Coll, C.G. 2009. "Racism and Child Health: A review of the literature and future directions." *Journal of Developmental and Behavioral Pediatrics* 30: 255.

Paradies, Y., Ben, J., Denson, N., Elias, A., Priest, N., Pieterse, A. 2015. "Racism as a Determinant of Health: A systematic review and meta-analysis." *PLoS One* 10, e0138511.

Post, Robert. 2012. "Interview with Robert Post." In *The Content and Context of Hate Speech*. Edited by M. Herz and P. Molnar, 11–36. Cambridge: Cambridge University Press.

Quillian, L., Pager, D., Hexel, O. and Midtbøen, A.H. 2017. "Meta-analysis of field experiments shows no change in racial discrimination in hiring over time." *Proceedings of the National Academy of Sciences* 114: 10870–10875.

Ren, X.S., Amick, B., and Williams, D.R. 1999. "Racial/ethnic Disparities in Health: The interplay between discrimination and socioeconomic status. *Ethnicity & Disease* 9: 151165.

Schneider, V. 2018. "Racism Knocking at the Door: The Use of Criminal Background Checks in Rental Housing." *U. Rich. L. Rev.* 53: 923.

Sentas, Vicki. 2018. "Beyond Media Discourse: Locating Race and Racism in Criminal Justice Systems." In *Media, Crime and Racism*, edited by Monish Bhatia, Scott Poynting, and Waqas Tufail, 359–79. Cham: Springer International Publishing.

Strossen, Nadine. 2018. *Hate: Why we should resist it with free speech, not censorship*. Oxford: Oxford University Press.

U.S. Department of Health and Human Services. 2001. *Mental Health: Culture, Race, and Ethnicity – A Supplement to Mental Health: A Report of the Surgeon General.* Rockville, MD: U.S. Department of Health and Human Services.

Wade, R., Shea, J.A., Rubin, D. and Wood, J. 2014. "Adverse Childhood Experiences of Low-Income Urban Youth." *Pediatrics* 134: e13–e20.

Waldron, Jeremy. 2012. *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press.

West, Caroline. 2012. "Words that Silence? Freedom of expression and racist hate speech." In *Speech and Harm: Controversies over free speech*. Edited by I. Maitra and M.K. McGowan, 222–248. Oxford: Oxford University Press.

Williams, D.R. 2000. "Race, Stress, and Mental Health." In *Minority Health in America*. Edited by C. Hogue, M. Hargraves, and K. Scott-Collins, 209–243. Baltimore: Johns Hopkins University Press.

Williams, D.R., Neighbors, H.W., and Jackson, J.S. 2003. "Racial/ethnic Discrimination and Health: Findings from community studies." *American Journal of Public Health* 93: 200–208.

Journal of Cognition and Neuroethics

Can, and Should, We Morally Enhance Psychopathic Individuals?

Ho Man Him
Maastricht University
Danish Research Center for Magnetic Resonance

Biography

Ho Man Him is a forensic psychology graduate student at Maastricht University, the Netherlands, and the Danish Research Center for Magnetic Resonance, Denmark. He will soon be pursuing a clinical psychology (forensic track) Ph.D. at Simon Fraser University, Canada. His interests include the assessment and treatment of psychopathy, underlying empathic and moral deficits, and associated psycho-legal issues.

Publication Details

Journal of Cognition and Neuroethics (ISSN: 2166-5087). February, 2023. Volume 9, Issue 1.

Citation

Ho, Man Him. 2023. "Can, and Should, We Morally Enhance Psychopathic Individuals?" *Journal of Cognition and Neuroethics* 9 (1): 35–49.

Can, and Should, We Morally Enhance Psychopathic Individuals?

Ho Man Him

Abstract

Whether or not the aim of treatment for those with psychopathy is to reduce criminality, or to fundamentally change them to fit within societal norms is debated, as well as the morality of associated "enhancement." This review covers contemporary literature and debates on moral enhancements, impairment, and the treatability of psychopaths across neuroethics and forensic psychology. I argue that by moral enhancement of psychopaths, we should mean "moral treatment of psychopaths," and that certain types of psychotherapy might be used to treat psychopaths, against the myth that they are untreatable. Moreover, I argue that the discussion should be focused on what is meant by "moral" and "enhancement (treatment)," with particular consideration of the distinction between passive/active and biomedical/traditional moral enhancement (treatment). Moreover, I caution how the ethics of moral enhancement hinges on associated changes in a psychopath's personality identity, who would benefit from the treatment, reversibility, and presence of safeguards.

Keywords

Psychopathy, Moral Enhancement, Biomedical Treatment, Narrative Identity, Forensic Psychology

> "Is it better for a man to have chosen evil than to have good imposed upon him?"
>
> —Anthony Burgess, *A Clockwork Orange*

In the dystopian novel *A Clockwork Orange*, Alex was a violent gang leader. After being convicted for murder, he voluntarily signed up for an aversion therapy that claimed to rehabilitate criminals across 2 weeks by pairing images of violence with fear, nausea, and paralysis-inducing drugs with Beethoven's 9th symphony in the background. This technique "programmed" Alex to only choose to be good by conditioning, and to be unable to resort to any form of violence, even when required to. He also exhibited side effects of averseness when listening to his favourite composer Beethoven, which eventually compelled him to attempt suicide to relieve himself from the pain.

The title *A Clockwork Orange* aptly highlights the novel's central thesis: if one is stripped of the freedom to choose between the morally good and bad, then they are not a human; they are clockwork/machine. Whether moral enhancement is ethical

has been a hotly debated topic, but as the landmark review by Specker et al. (2014) suggests: we need to shift our focus of moral enhancement to those with pathological deficiencies. Hence, I would like to focus on psychopathic individuals. Despite not being an official clinical diagnosis in the DSM-5, psychopathy is a robust disorder characterized by moral impairments that provides useful insights, especially within forensic contexts (Hare and Neumann 2009). Before we discuss the possibilities of moral enhancement in psychopaths, it would be imperative to discuss what psychopathy and moral enhancement actually entail and why it might be important to enhance morals. I will review the existing literature, and offer a critical analysis of the issue.

Dangers of Psychopathy and the Catch-22 Dilemma

The term "psychopath" is often associated with charismatic serial killers or chronic criminal offenders. Some would consider the term synonymous with violence, and the disorder untreatable (Skeem et al. 2011). This is a sentiment that is shared by many clinicians, who also believe psychopaths cannot be cured (Salekin 2002), and after release, could offend more than other "types" of offenders (Rice et al. 1992). This has led to unfortunates scenarios in some forensic institutions: because psychopaths do not respond well to treatment, they should not take part in treatment. However, because of this, they are unable to leave the institutions via release or parole, thus resulting in a Catch-22 where there is no escape. Another argument for the inability to treat psychopaths centres around biomedical enhancement, namely that psychopaths cannot and should not be cured with neuromodulatory drugs, as this can change one's social and moral outlook, and hence, can alter their identity radically (Maibom 2014).

What is a Psychopath? And how are they "Morally Impaired?"

The definition of psychopathy has been changing continuously across time, but the 16 Diagnostic Criteria set out by Cleckley (1976) in "The Mask of Sanity" have been the most influential operational definition (Patrick 2018, 5). Many of the symptoms described (e.g., lacking remorse/shame, untruthfulness and insincerity) are moral impairments that characterize the disorder. The same applies for Hare's Psychopathy Checklist-Revised (PCL-R), which has a significant focus on criminal behaviour and violence that comes as a consequence of moral impairments (e.g., Pathological Lying, Callousness and Lack of Empathy). However, there is considerable debate on whether

the PCL-R is a good measure of psychopathy. For instance, criminal behaviour might be a correlate rather than central construct of psychopathy (Skeem and Cooke 2010). To integrate conflicting literature, Patrick, Fowles, and Krueger (2009) created the triarchic model of psychopathy. Specifically, the model stipulates that psychopathy consists of three distinct, but overlapping phenotypic constructs: disinhibition, boldness, and meanness, where all of them have to be present for psychopathy. The key moral impairment would be meanness, where disinhibition acts as a catalyst. This theory also elegantly ties in evidence from cognitive psychology and neuroscience. That is, Blair's Integrated Emotional Systems highlights how meanness could develop from dysfunctional emotional reactivity (where individuals are not sensitive to distress cues, and hence moral/social transgressions occur), owing to deficits in the amygdala. Newman's Response Modulation Theory could explain how cognitive attentional-deficits could impair inhibitory control and punishment learning (Patrick 2022). One could then posit that an effective enhancement/intervention would need to tackle (one of) these three phenotypic traits.

A main reason for moral enhancement is because psychopathy is costly to society. Kiehl and Hoffman (2011) estimated that offending psychopaths cost the U.S. 460 billion per year in criminal social costs, without accounting for psychological costs of victims. It is a risk factor for violence, and there are high costs for non-treatment (Hare 1999). Given the economic, social, and psychological costs of psychopathy (notably violence), if there was a pill that would reduce violent tendencies and immoral behaviour of psychopaths, why shouldn't we instantly use it?

What is Moral Enhancement? Should Psychopaths be Enhanced?

Shook (2012) defines moral enhancement as the modification of brain processes to produce more moral conduct, and to make one more likely to do the morally right thing. He also reminds us that only if increasing moral motivation means greater moral conduct, can it be considered a reliable method of enhancing morality, as illustrated by Douglas (2008). Simkulet (2012) however, states that Shook's definition is flawed. Forcing agents to act rightly by preventing/making it difficult for them from acting wrongly should not be considered enhancement but compulsion, meaning they are forced against their own will to act morally (e.g., *A Clockwork Orange*). Simkulet (2012) posits that moral enhancement facilitates usage of one's free will to make people more likely to succeed in their goals, which is what separates moral from other types of enhancements. In other

words, the distinction it is not as simple as Shook's idea of "environing social contexts" (2012, 3). Focquaert and Schermer (2015) further clarify that moral enhancement is the strengthening of moral capacities, leading to cognitive, affective, and motivational changes in moral decision making and behaviour. A long-term and stable enhancement should entail understanding of what differentiates morally right from wrong, to distinguish enhancement from mere behavioural control, involving responsiveness to moral reasons. There is a general lack of consensus on what constitutes morals, namely how "moral development" depends on which ethical system or theory one agrees with and moral pluralism dictates that there can be (conflicting) moral views that are considered equal and respected. In fact, Brooks (2012) goes as far as saying enhancement itself might violate the equality of reasonable pluralism. Moreover, Specker et al. (2014) aptly summarized that what counts as improvements depends on contexts and roles (e.g., we need detached surgeons to remove brain tumors, but not detached mothers to raise children). However, as Kahane and Savulescu (2013) would be quick to point out, this might call for precision, and not elimination of enhancement. Moreover, one could argue there is significant overlap of morals across different ethical systems (e.g., altruism, fairness, and empathy [Persson & Savulescu, 2013]). This corresponds with what Shook (2012) called Minimal Moral Commonsensism, where he argues we should enhance commonly accepted views of morality, enhancing at least one of the following moral contexts: (1) Appreciation; (2) Decisions; (3) Judgements; (4) Intentions; and (5) Willpower. Kabasenche (2012) encourages us to think of Shook's concept as a Moral Quotient (MQ) with, for example, increasing moral appreciation increasing the MQ score. He then argues that with the absence of other moral contexts, one is not truly moral, and we can only measure morals holistically. This is a fair critique because enhancement of moral appreciation doesn't necessarily mean moral action, despite an unfair assumption of MQ being measured linearly (e.g., each moral context could be weighted differently or be interconnected). But the question still remains: how do we define what is "moral?"

I believe Lev (2012) provides an interesting alternative (i.e., moral enhancement should focus on basic moral capacities that enable exercise of moral agency). Namely, he suggests: (1) Critical Reflection; (2) Impartiality; (3) Imaginative; and (4) Interpretative Abilities. This could be reconciled with Simkulet's proposal, as it increases the likelihood that one could exercise their free will in achieving their goals. However, in the context of psychopathy, I am sceptical as to whether Lev's proposal might be feasible. Some psychopaths are perfectly capable of cognitive empathy, or in understanding what is morally right from wrong (Cima et al. 2010). Their main impairment isn't in

understanding morals but not caring about such knowledge and its consequences, or that they are unable to automatically access this knowledge when engaging in goal-directed behaviour (e.g., due to attentional deficits) (Drayton et al., 2017; Vitale et al., 2016). In fact, Cleckley (1976) himself observed psychopaths show no evidence of a deficit in complex matters of judgement, as long as they are not direct participants. Or in the words of our protagonist Alex, "I see what is right and approve, but I do what is wrong." Moreover, as Horstkötter et al. (2012) argues, one should distinguish between treatment and enhancement. Nick Bostrum defines enhancement as an elevation beyond normal levels (Bostrum 2008), while Dorothee Horstkötter points out that those with pathological moral/antisocial impairments that deviates from the norm (e.g., psychopaths) need medical treatment to reduce such impairments, and enhancement is not required within this context. At least in the context of psychopaths, the appropriate term would be "treatment," not "enhancement." This raises several questions, namely: What is considered "normal?" When is moral functioning pathological? For example, lawyers, hedge fund managers, or world leaders are sometimes considered "successful psychopaths," or as Hare and Babiak (2006) would call them: "Snakes in Suits." Should they be enhanced/treated as well? This further leads to a fundamental issue: Who decides what is morally better? The morality of a superior moral agent in control is a different debate, but regardless, there should be safeguards to avoid abuse of power and usage of enhancements that are irreversible, continuously reviewed, and revised.

Feasibility of Treating Psychopaths

Going back to our introduction, if psychopaths were truly "untreatable," then our ethical debate would only be a mildly stimulating thought experiment. Is this really true (e.g., in forensic settings)? D'Silva (2004) systematically reviewed 24 psychopathy intervention studies, and found that no study met the standard for an acceptable study to answer whether "Treatment[s] make psychopaths worse." Notwithstanding the severe methodological flaws, they concluded that the PCL-R and treatment response association is still inconsistent. In a recent review, De Ruiter and Hildebrand (2022) found that psychopathy is not untreatable, and in fact works especially well if it is personalised and continued over long durations. They point out the myth that psychopaths cannot engage in therapeutic alliance, and how there is no evidence high scoring psychopaths seek treatment to manipulate others.

Moreover, the authors cite the general effectiveness of Cognitive Behavioural Therapy (CBT) regardless of PCL-R scores, and the potential for Schema Therapy (ST) specific for psychopaths. They especially mention a case study which showed how ST could successfully treat a PCL-R psychopath over 4 years, changing both the affective and interpersonal facets (closely linked to moral deficits) without fundamentally altering his identity (Chakhssi et al. 2014). Of course, one would rightfully criticize that overarching conclusions should not be made from individual cases (Crockett et al. 2014). However, in a recent randomized control trial (RCT), ST was found to be more effective than regular treatments for forensic populations in enhancing rehabilitation, and reducing personality disorder symptoms, including those with antisocial and borderline traits (Bernstein et al., 2021). This suggests a potential to treat violent offenders, to make them understand and behave better morally, including those with psychopathy. Returning to Patrick's triarchic model, ST helps one meet their own emotional needs by identifying patterns of negative thinking and developing new coping mechanisms. Thus, I would argue this mainly acts in reducing the disinhibition facet (e.g., understanding and evaluating consequences of actions and thoughts). ST would also fulfil the earlier definitions of moral enhancement, providing a true understanding of morals paired with corresponding action within appropriate contexts. By working on disinhibition, this alleviates the problem that psychopaths are unable to access/apply morals they understand. I would agree with the concern of Specker et al. (2014) that some researchers overestimate feasibility of moral enhancement (e.g., genetic modification of vices and virtues, using Deep Brain Stimulation [DBS] to target phenotypic traits characteristics, etc.); but, in the case of (offender/PCL-R) psychopathy, there is reason to believe that moral treatment is indeed feasible, at least by means of cognitive therapies, so they will no longer be trapped in Catch-22s.

Differences between Biomedical and Traditional Moral Enhancements of Psychopaths

The feasibility of treating psychopaths might not apply to all moral enhancements. One might need to distinguish between biomedical (e.g., drugs, tDCS) and traditional forms (e.g., ST, moral education) of enhancement. Glannon (2014) points out the lack of empirical studies showing the effectiveness of psychotropic medication in reducing/eliminating psychopathic traits/behaviour, while Hübner and White (2016) warns us of the ethical flaws in using DBS for treating psychopathy (i.e., because there is no

individual medical benefit, and voluntary informed consent). More crucially, those who benefit most from moral (bio)enhancement might be society (e.g., safety), and not the individual offender. Moreover, our knowledge is limited regarding side effects of moral enhancements, especially biomedical forms. Highly invasive procedures such as DBS require brain surgery, and have potential side effects. For example, DBS in Parkinson's Disease could lead to cognitive, behavioural or psychiatric side effects, despite being reversible (Clausen 2010). At present, we do not know whether the same applies for psychopaths, as it did for our protagonist Alex. This would especially be risky for procedures that are non-reversible, as Specker et al. (2014) illustrates using stem-cell injections. This constitutes a key difference between biomedical and traditional forms of enhancement: there is potential for more side effects, and also irreversibility for some biomedical techniques compared to traditional treatment. A larger problem for moral enhancing treatment of psychopaths as posed by Maibom (2014) is that since psychopathy is a personality disorder, to treat it would be to change one's identity drastically.

Perhaps Macbeth best illustrated this (Shakespeare 1992/1606, 46 - 47):

> I dare do all that become a man;
>
> Who dares do more is none.

Shakespeare reminds us despite Macbeth's initial reluctance, by daring to kill Duncan, he dared to do more, and the more he dared the less human he became. This illustrates a main concern of moral enhancement (i.e., once we dare to accept and actively modify ourselves, when will we eventually lose our sense of humanity?). Focquaert and Schermer (2015) acknowledge this, warning of the dangers corresponding with changes in narrative identity. Narrative identity consists of central and salient characteristics that build a person's identity. When one's narrative identity changes, this should be incorporated without compromising the sense of self for the continuity of narrative. The authors give an example of how moral enhancement could cause abrupt or concealed identity changes that are disruptive. If after a moral enhancement treatment, a psychopath now suddenly becomes warm and empathetic, this could threaten the continuity of their narrative identity. The authors further posit identity changes that could be unnoticed by the treated patient, but eventually threatens the autonomy of the self with the associated incoherence.

Whether or not specific treatments should be used to morally enhanced psychopaths might be evaluated using Focquaert and Schermer's (2015) classification of treatments.

They proposed a distinction between active v passive; direct v indirect interventions. Direct interventions *target the brain* in order to change thoughts and behaviour, while indirect interventions *change thoughts patterns and behaviour* to rewire one's brain structure and functioning. This is essentially a distinction between biomedical and traditional forms of interventions.

The second distinction is more interesting: active treatments require specific psychological/behavioural efforts from the individual to reach a desired end, while passive interventions do not need this. In other words, active interventions are done *with* participants, while passive ones are done *to* them. The authors argue passive interventions are more dangerous ethically, as this might compromise a person's autonomy and identity. Participants are unable to withdraw consent during such treatment, which can lead to sudden/concealed narrative identity changes. In contrast, indirect interventions do not have this problem, as individuals are involved continuously (e.g., in ST, the individual has to actively identify their own negative patterns). They do acknowledge the potential problems of direct neuromodulations, which is more likely to be passive (i.e. the device does everything, more likely to bypass conscious reflection, deliberation, and choice. However, they also suggest that direct interventions could be justified, if there are safeguards; e.g., proper informed consent, procedures, and pre-post-intervention counselling). This means that an individual has made a choice freely and have an active role, with corresponding insight and reflection, to incorporate passively induced changes into their narrative identity. Considering safeguards, a deeper cost-benefit understanding of enhancement techniques, direct and indirect interventions could be equally justified ethically.

Coercion

One could challenge whether psychopathic offenders freely choose treatment? If a treatment (direct or indirect) is shown to be effective, is it justified to enforce it? Indirect treatments are less of a problem here, as they simply would not work without therapeutic alliance, and the direct involvement of the individual in question.

The debate on whether psychopaths should be forced to morally bio-enhance concentrates on the violation of freedom of thought (i.e., is the State justified in intervening forcefully or are they violating an offender's freedom?). Craig (2016) is strongly against the intrusion of a psychopath's freedom. He argues that there is a fundamental right to mental integrity, which should be protected to prevent disruption

of narrative identity, and hence autonomous human agency. Peterson and Kragh (2017) however, respond by arguing that being confined to prison is also associated with a loss of freedom of thought, e.g. it could lead to inability to initiate activity, chronic depression, or loss of sense of reality (32). It is not clear whether bioenhancement is any more damaging than incarceration. Allowing forced imprisonment but not forced bioenhancement rehabilitation would be inconsistent, and a double standard.

Curtis (2012) would support forced enhancement in general, arguing that everything boils down to classifications of enhancements. For instance, supplemental enhancements have a less severe impact than strengthening ones, while emotional enhancements have more impact than cognitive or volitional ones. In fact, if the enhancement was safe and effective, prevents harm to others, and reintegrates one into society, this can be more cost-effective than incarceration, if we apply a utilitarian argument. However, one might also argue from a deontological perspective whether sacrificing human autonomy is ever justified, or to quote the Chaplain from *A Clockwork Orange*: "Goodness is something chosen. When a man cannot choose, he ceases to be a man." Fundamentally, if coercion is ever used, then it should only be used as a means of last resort, and not as commonplace, in order to respect the freedom and autonomy of the individual as much as possible (Nedopil 2016).

More recently, Baccarini and Malatesti (2017) proposed an open justification to treating psychopathy using moral bioenhancement. They say one should only prescribe what they would also prescribe to others, and they believe psychopaths would want other psychopaths to be morally bioenhanced. I believe this argument is flawed. From a practical standpoint, it is not necessary to use bioenhancement as there are better validated alternatives (e.g., Schema Therapy). Moreover, as described above, psychopaths are morally impaired, but show little deficits in cognitive empathy or rational understanding. There is little reason to believe that they lack the volition to make rational decisions according to their system of reasons. Moreover, from a neuroethics standpoint, as Sirgiovanni and Garasic (2020) state, there is evidence that "the psychopath's cognitive-affective system would consistently justify reasons against mandatory moral bioenhancement to herself, even if she wishes differently for others, and that the prescription cannot be extended" (2). Adding the problems of irreversibility and radical changes in narrative identity of bioenhancement to the practical, empirical, and neuroethical challenges posed, the open justification argument might be limited.

Conclusions

Alex's story ends with him being reverted to his "normal" violent self, having learnt nothing from his experiences, in contrast to us. I argued that whether psychopaths should be morally enhanced depends on the definition and measurement of "moral" and whether it fits with existing knowledge of psychopathy. I also pointed out by "moral enhancement," what we really mean is "treatment" for psychopaths. Moreover, I explored the myth of how "psychopaths are untreatable," and that there are currently effective means to do so. Lastly, I explained the distinctions between biomedical and traditional forms of moral enhancement for psychopaths could lie in who it benefits, associated changes in identity, our knowledge of side effects, irreversibility, active/passiveness, and presence of safeguards.

Future Research

Considering the contents of this review, this leaves us with some ideas for future research:

(1) What is the best method for moral enhancement in psychopaths? Is it better to use biomedical techniques in conjunction with or separately from traditional methods (cf. Kabasenche 2012)?

(2) How do we design safeguards to prevent moral enhancement methods from falling into the wrong hands, e.g., state-control, or psychopaths who strive to cause ultimate harm (cf. Tonkens 2012)?

(3) Is it reasonable to force children with psychopathic traits (e.g. who commit violent acts) coercively to use moral enhancement/interventions, and would it be stigmatizing and a self-fulfilling prophecy that results in moral decline (cf Horstkötter et al. 2012; Glannon 2014)?

These are only some of the suggestions for future research. I hope that this review presented a comprehensive picture of the literature, connecting arguments from neuroethics with a clinical, empirical understanding of psychopathy for further research questions to be raised.

References

Baccarini, Elvio, and Luca Malatesti. 2017. "The Moral Bioenhancement of Psychopaths." *Journal of Medical Ethics* 43 (10): 697–701. https://doi.org/10.1136/medethics-2016-103537.

Bernstein, David P., Marije Keulen-de Vos, Maartje Clercx, Vivienne de Vogel, Gertruda C. M. Kersten, Marike Lancel, Philip P. Jonkers, et al. 2023. "Schema Therapy for Violent PD Offenders: A Randomized Clinical Trial." *Psychological Medicine* 53 (1): 88–102. https://doi.org/10.1017/S0033291721001161.

Blair, R. J. R. 2005. "Applying a Cognitive Neuroscience Perspective to the Disorder of Psychopathy." *Development and Psychopathology* 17 (3): 865–91. https://doi.org/10.1017/S0954579405050418.

Brooks, Thom. 2012. "Moral Frankensteins." *AJOB Neuroscience* 3 (4): 28–30. https://doi.org/10.1080/21507740.2012.721467.

Chakhssi, Farid, Truus Kersten, Corine de Ruiter, and David P. Bernstein. 2014. "Treating the Untreatable: A Single Case Study of a Psychopathic Inpatient Treated with Schema Therapy." *Psychotherapy* 51: 447–61. https://doi.org/10.1037/a0035773.

Christen, Markus, and Darcia Narvaez. 2012. "Moral Development in Early Childhood Is Key for Moral Enhancement." *AJOB Neuroscience* 3 (4): 25–26. https://doi.org/10.1080/21507740.2012.721460.

Cima, Maaike, Franca Tonnaer, and Marc D. Hauser. 2010. "Psychopaths Know Right from Wrong but Don." Care." *Social Cognitive and Affective Neuroscience* 5 (1): 59–67. https://doi.org/10.1093/scan/nsp051.

Clausen, Jens. 2010. "Ethical Brain Stimulation – Neuroethics of Deep Brain Stimulation in Research and Clinical Practice." *European Journal of Neuroscience* 32 (7): 1152–62. https://doi.org/10.1111/j.1460-9568.2010.07421.x.

Cleckley, Hervey M. 1976. *The Mask of Sanity: An Attempt to Clarify Some Issues about the so-Called Psychopathic Personality*. Saint Louis: The C. V. Mosby Company.

Craig, Jared N. 2016. "Incarceration, Direct Brain Intervention, and the Right to Mental Integrity – a Reply to Thomas Douglas." *Neuroethics* 9 (2): 107–18. https://doi.org/10.1007/s12152-016-9255-x.

Crockett, Molly J. 2014. "Moral Bioenhancement: A Neuroscientific Perspective." *Journal of Medical Ethics* 40 (6): 370–71. https://doi.org/10.1136/medethics-2012-101096.

Curtis, Benjamin L. 2012. "Moral Enhancement as Rehabilitation?."*AJOB Neuroscience* 3 (4): 23–24. https://doi.org/10.1080/21507740.2012.721448.

Douglas, Thomas. 2008. "Moral Enhancement." *Journal of Applied Philosophy* 25 (3): 228–45. https://doi.org/10.1111/j.1468-5930.2008.00412.x.

D."ilva, Karen, Conor Duggan, and Lucy McCarthy. 2004. "Does Treatment Really Make Psychopaths Worse? A Review of the Evidence." *Journal of Personality Disorders* 18 (2): 163–77. https://doi.org/10.1521/pedi.18.2.163.32775.

Dutton, Kevin. 2012. *The Wisdom of Psychopaths: What Saints, Spies, and Serial Killers Can Teach Us About Success*. New York: Scientific American/Farrar, Straus and Giroux.

Focquaert, Farah, and Maartje Schermer. 2015. "Moral Enhancement: Do Means Matter Morally?."*Neuroethics* 8 (2): 139–51. https://doi.org/10.1007/s12152-015-9230-y.

Glannon, Walter. 2014. "Intervening in the Psychopath." Brain." *Theoretical Medicine and Bioethics* 35 (1): 43–57. https://doi.org/10.1007/s11017-013-9275-z.

Hare, Robert D. 1999. "Psychopathy as a Risk Factor for Violence." *Psychiatric Quarterly* 70 (3): 181–97. https://doi.org/10.1023/A:1022094925150.

Hare, Robert D, and Craig S Neumann. 2009. "Psychopathy: Assessment and Forensic Implications." *The Canadian Journal of Psychiatry* 54 (12): 791–802. https://doi.org/10.1177/070674370905401202.

Horstkötter, Dorothee, Ron Berghmans, and Guido de Wert. 2012. "Moral Enhancement for Antisocial Behavior? An Uneasy Relationship." *AJOB Neuroscience* 3 (4): 26–28. https://doi.org/10.1080/21507740.2012.721451.

Hübner, Dietmar, and Lucie White. 2016. "Neurosurgery for Psychopaths? An Ethical Analysis." *AJOB Neuroscience* 7 (3): 140–49. https://doi.org/10.1080/21507740.2016.1218376.

Kabasenche, William Paul. 2012. "Moral Enhancement Worth Having: Thinking Holistically." *AJOB Neuroscience* 3 (4): 18–20. https://doi.org/10.1080/21507740.2012.721464.

Kahane, Guy, and Julian Savulescu. 2015. "Normal Human Variation: Refocussing the Enhancement Debate." *Bioethics* 29 (2): 133–43. https://doi.org/10.1111/bioe.12045.

Kiehl, Kent A., and Morris B. Hoffman. 2011. "The Criminal Psychopath: History, Neuroscience, Treatment, and Economics." *Jurimetrics* 51: 355–97.

Lev, Ori. 2012. "Enhancing the Capacity for Moral Agency." *AJOB Neuroscience* 3 (4): 20–22. https://doi.org/10.1080/21507740.2012.721462.

Maibom, Heidi L. 2014. "To Treat a Psychopath." *Theoretical Medicine and Bioethics* 35 (1): 31–42. https://doi.org/10.1007/s11017-014-9281-9.

Nedopil, Norbert. 2016. "Special Considerations in Forensic Psychiatry." In *The Use of Coercive Measures in Forensic Psychiatric Care: Legal, Ethical and Practical Challenges*, edited by Birgit Völlm and Norbert Nedopil, 135–49. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-26748-7_8.

Patrick, Christopher J. 2022. "Psychopathy: Current Knowledge and Future Directions." *Annual Review of Clinical Psychology* 18 (1): 387–415. https://doi.org/10.1146/annurev-clinpsy-072720-012851.

Patrick, Christopher. J. 2018. *Handbook of psychopathy (2nd ed)*. The Guilford Press.

Patrick, Christopher J., Don C. Fowles, and Robert F. Krueger. 2009. "Triarchic Conceptualization of Psychopathy: Developmental Origins of Disinhibition, Boldness, and Meanness." *Development and Psychopathology* 21 (3): 913–38. https://doi.org/10.1017/S0954579409000492.

Persson, Ingmar, and Julian Savulescu. 2013. "Getting Moral Enhancement Right: The Desirability of Moral Bioenhancement." *Bioethics* 27 (3): 124–31. https://doi.org/10.1111/j.1467-8519.2011.01907.x.

Petersen, Thomas Søbirk, and Kristian 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.

Rice, Marnie E., Grant T. Harris, and Catherine A. Cormier. 1992. "An Evaluation of a Maximum Security Therapeutic Community for Psychopaths and Other Mentally Disordered Offenders." *Law and Human Behavior* 16 (4): 399–412. https://doi.org/10.1007/BF02352266.

Ruiter, Corine de, and Martin Hildebrand. 2022. "Therapeutic Considerations and Interventions for Psychopathy." In *The Complexity of Psychopathy*, edited by Jennifer E. Vitale, 359–80. Dangerous Behavior in Clinical and Forensic Psychology. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-83156-1_14.

Salekin, Randall T. 2002. "Psychopathy and therapeutic pessimism. Clinical lore or clinical reality?." *Clinical psychology review vol*. 22,1 (2002): 79-112. https://doi.org/10.1016/s0272-7358(01)00083-6.

Shakespeare, William. 1992. *Macbeth*. Wordsworth Classics.

Shook, John R. 2012. "Neuroethics and the Possible Types of Moral Enhancement." *AJOB Neuroscience* 3 (4): 3–14. https://doi.org/10.1080/21507740.2012.712602.

Simkulet, William. 2012. "On Moral Enhancement." *AJOB Neuroscience* 3 (4): 17–18. https://doi.org/10.1080/21507740.2012.721449.

Sirgiovanni, Elisabetta, and Mirko Daniel Garasic. 2020. "Commentary: The Moral Bioenhancement of Psychopaths." *Frontiers in Psychology* 10. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02880.

Skeem, Jennifer L., and David J. Cooke. 2010. "Is Criminal Behavior a Central Component of Psychopathy? Conceptual Directions for Resolving the Debate." *Psychological Assessment* 22: 433–45. https://doi.org/10.1037/a0008512.

Skeem, Jennifer L., Devon L. L. Polaschek, Christopher J. Patrick, and Scott O. Lilienfeld. 2011. "Psychopathic Personality: Bridging the Gap Between Scientific Evidence and Public Policy." *Psychological Science in the Public Interest* 12 (3): 95–162. https://doi.org/10.1177/1529100611426706.

Specker, Jona, Farah Focquaert, Kasper Raus, Sigrid Sterckx, and Maartje Schermer. 2014. "The Ethical Desirability of Moral Bioenhancement: A Review of Reasons." *BMC Medical Ethics* 15 (1): 67. https://doi.org/10.1186/1472-6939-15-67.

Tonkens, Ryan. 2013. "Feeling Good About the End: Adderall and Moral Enhancement." *AJOB Neuroscience* 4 (1): 15–16. https://doi.org/10.1080/21507740.2012.757567.

Vitale, Jennifer. E., Baskin-Sommers, Arielle. R., Wallace, John. F., Schmitt, W. A., & Newman, Joseph. P. 2016. "Experimental investigations of information processing deficiencies in psychopathic individuals: Implications for diagnosis and treatment." In Gacono, Carl, B. (Ed.), *The clinical and forensic assessment of psychopathy: A practitioner." guide*, 54–72. Routledge/Taylor & Francis Group.

Waller, Rebecca, and Luke Hyde. 2017. "Callous-Unemotional Behaviors in Early Childhood: Measurement, Meaning, and the Influence of Parenting." Child development perspectives vol. 11,2: 120-126. https://doi.org/10.1111/cdep.12222.

Journal of Cognition and Neuroethics

Conceptual and Empirical Pinpointing of Consciousness

Tobias A. Wagner-Altendorf
Northwestern University
University of Lübeck
Munich School of Philosophy

Biography
Tobias A. Wagner-Altendorf is a clinical neurologist at the University of Lübeck, Germany. From 2022 to 2023, he was visiting scholar working in cognitive neuroscience at Northwestern University in Evanston, Illinois. Also, he's pursuing a PhD in the philosophy of mind at Munich School of Philosophy in Munich, Germany.

Acknowledgements
I am grateful for the lively discussions on empirical consciousness research during the "Mind and Brain" course held by Dr. Ken Paller at Northwestern University in Spring Quarter 2022.

Conflict of Interest
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding
The work was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GEPRIS 465881133.

Publication Details
Journal of Cognition and Neuroethics (ISSN: 2166-5087). August, 2023. Volume 9, Issue 1.

Citation
Wagner-Altendorf, Tobias A. 2023. "Conceptual and Empirical Pinpointing of Consciousness." *Journal of Cognition and Neuroethics* 9 (1): 51–65.

Conceptual and Empirical Pinpointing of Consciousness

Tobias A. Wagner-Altendorf

Abstract
Consciousness is targeted by both philosophers and neuroscientists; but different methodological premises and even different conceptions about what conscious experience is and how the challenges and potential problems associated with consciousness research should be formulated underlie the different approaches. Namely, whereas empirical data and the constant refinement of experimental procedures to expand and modify this body of empirical data and resulting empirical theories are crucial to neuroscience, the significance of empirical knowledge to philosophy is less clear: Although empirical data certainly can influence philosophical concepts, the latter are nonetheless prerequisites of empirical research itself and thus may themselves not be empirically testable. The present paper elaborates from a multidisciplinary, neuroscientist-philosopher's perspective the relation of philosophical concepts and empirical research on consciousness, drawing on two exemplary controversies from the philosophy of mind – on the ontological status of experiential properties and on free will. Consequences from both the scientific and the philosophical standpoint are discussed.

Keywords
Neuroscience, Philosophy of Mind, Hard Problem of Consciousness, Free Will

Introduction

Consciousness – although in the past sometimes accused of being ignored by the sciences – is targeted by several disciplines including philosophy and neuroscience. However, besides differing methodologies, philosophy and neuroscience might not even agree on the adequate definition of conscious experience as well as the adequate description of the challenges and potential problems associated with consciousness research.

As consciousness – albeit we're intimately familiar with it, more familiar probably than with every other phenomenon – has proven to be quite challenging to sufficiently define, researchers often offer an intuitive way to circumscribe it: consciousness is "what vanishes every night when we fall into dreamless sleep and reappears when we wake up or when we dream" (Tononi 2008). Ned Block, on defining (qualitative) consciousness, famously quoted – "only half in jest" – a Louis Armstrong dictum on jazz: "If you got

to ask, you ain't never gonna get to know" (Block 1978). Philosophers have used to distinguish *qualitative* or *phenomenal consciousness* from *psychological* or *access consciousness* (e.g., Block 1995, and Chalmers 1996) – with the former presumably posing the greater difficulties for, e.g., a reductionist account of the mental –, but one might question the idea of a strict dichotomy between intentional and phenomenal aspects of the mental, as every intentional reference, and ultimately every (also non-iconic) thinking, might be ascribed qualitative character (e.g., Siewert 1998). Following an intuitive definition approach, having consciousness may for the present purpose be defined as *subjectively experiencing the world from a first-person perspective*.

More generally speaking, philosophy is concerned with *conceptual* questions of consciousness: e.g., how does the concept or the notion of the mental fit with the concept or the notion of the physical? How are those two interrelated, and can one concept eventually be reduced to the other? And, to add a meta-level to the discussion: what means do we have and should we have to decide about those questions?

Another debate pointing to the conceptual nature of philosophical controversies on consciousness is the discussion of free will and agency. What is the adequate concept of free will – of a decision or of an action appropriately characterized as free –, and what does this concept presuppose about the nature or the make-up of our world? One might argue that the adequate conception of free will requires determinism, that it requires indeterminism, or that it is agnostic with respect to whether it requires determinism or indeterminism. (One might also argue that the notion of free will is inherently inconsistent in the settings of either determinism or indeterminism, and that thus there is no adequate conception of free will.) In any case, *conceptual* issues are addressed, irrespective of empirical findings.

Au contraire, empirical data and the constant refinement of experimental procedures to expand and modify this body of empirical data and resulting empirical theories are crucial to neuroscience: empirical research on consciousness seeks to identify brain structures and networks associated with and *responsible for* conscious experience, and, e.g., distinguish them from regions and networks involved in attentional processes or processes related to the monitoring and reporting of conscious experience. Importantly, not only a mapping of anatomical structures implicated in consciousness drives empirical research, but theorizing on the *functional* organization and interplay between distinct brain circuits necessary and sufficient for consciousness, i.e., on the neurophysiology forming the foundation of our conscious experience.

The significance of empirical knowledge to the philosopher, however, is less clear: although empirical data certainly can influence philosophical concepts, the latter are

nonetheless prerequisites of empirical research itself and thus typically themselves not empirically testable.

The present paper aims at disentangling the – often confounded – neuroscientific, i.e., empirical, and philosophical, i.e., conceptual, approaches to consciousness. To elaborate how the two are related, yet distinct, I will be drawing on two exemplary controversies from the philosophy of mind – on the ontological status of experiential properties and on free will. Consequences from a multidisciplinary perspective incorporating both the scientific and the philosophical standpoint are discussed.

Consciousness, the Ontological Status of Experiential Properties, and Empirical Research

Consciousness, as stated in the introduction, is hard to define; and despite of us being immediately familiar with "conscious experience," it is probably not one single und homogenous phenomenon, but is having different yet related aspects. One common distinction, as noted, is to discriminate the functional aspects of consciousness from its qualitative aspects, or "access consciousness" from "phenomenal consciousness." As another term, roughly equating "phenomenal," philosophers have coined the notion of the "experiential," i.e., experience-related or experience-bearing properties (with experience understood as essentially qualitative, phenomenal experience).

The ontological status of experiential properties has been extensively discussed in the philosophical literature over the past decades: can their qualitative character eventually be reduced to the (non-qualitative) character of physical properties? Reductive physicalism emphatically endorses a positive answer, pointing to eventually-shown-to-be-reducible properties such as liquidity, heat and life, whereas non-reductive positions, namely dualism and non-reductive physicalism, deny the reducibility of the experiential, referring to the now-classical anti-reductionist arguments, e.g., by Nagel (1974), Jackson (1982), and Chalmers (1996).

My point here is not in debating reductionist vs. anti-reductionist arguments, but to relate this conceptual philosophical debate to (empirical) neuroscience. An extensive, and ever more rapidly increasing, body of empirical research has been built up, in search for and in pinpointing of what has been called "the neural correlate(s) of consciousness" (NCC) (Crick & Koch 1998). Importantly, the search not only for correlational, but for *causal* relations between (a specific type of) conscious experiencing and (a specific type of) brain activation pattern drives neuroscientific research: to identify the "minimum

neural mechanisms jointly *sufficient* for any one specific conscious experience" (Koch et al. 2016; my emphasis).

Empirical theories on (the neural correlates and causes of) consciousness have taken different stances on explaining consciousness on a (neuro-)biological basis, and on disentangling "true" NCCs from "mere" neural prerequisites and/or consequences of consciousness (see Seth & Bayne 2022, for a recent review). The most influential empirical theories of consciousness include global workspace theories – holding that an item becomes conscious through "information broadcasting" within a widespread neuronal "workspace" including parietal and prefrontal areas, with the P3b being the hallmark electrophysiological index of conscious access (e.g., Dehaene & Changeux 2011; Mashour et al. 2020) –, integrated information theories – holding that physical systems are the basis for consciousness precisely if they constitute a network combining functional specialization with functional integration, so that their integrated information is high (e.g., Tononi 2008; Tononi et al. 2016) –, and re-entry theories – holding that recurrent processing is the crucial neural ingredient of consciousness, and thus assuming that phenomenal consciousness is possibly much more widespread than our *access* to and later reporting on it (e.g., Lamme 2006 2010).

It is not the aim of the present paper to compare the different empirical theories of consciousness, but to, meta-empirically, contrast them to a philosophical approach to consciousness. So – how do neuroscience and empirical research relate to the *concepts* of consciousness in the philosophy of mind, and to the conceptual philosophical questions about, e.g., the ontological relation of the notions of the mental and the physical?

One might assume that neuroscientific theories philosophically favor a reductionist, i.e., a physicalist account of the mental over the non-reductionist account. And indeed, dualistic theories – being the prime example of a non-reductionist account of the mental – have been harshly attacked by some scientists, and attributed, e.g., a "misguided intuition" (Dehaene 2014, 2), as if they were at odds with neuroscience.

Obviously, however, the case is not as clear-cut. This can be illustrated considering particular philosophical positions. Think, to begin with, of the mental-to-physical reductionist view: *strictly speaking* – it's saying – *there is nothing mental as an entity of its own; the mental eventually reduces to the physical. (The physical world, in particular our brain, makes up the mind.)* And now consider the very *opposite* philosophical view – also reductionist, but in the physical-to-mental direction: *strictly speaking, there is nothing physical as an entity of its own; the physical eventually reduces to the mental. (The mind makes up the physical word, including our brain.)*

The two positions couldn't philosophically be more divergent – but it seems very hard to decide upon and between them based on empirical measures, and it is reasonable to assume that proponents of each position would consider their view to perfectly fit with what science has told us and will tell us.

Science provides us, says the mental-to-physical reductionist, with an increasingly detailed picture about how physical processes underlie mental processes, thus bolstering the idea that the physical is fundamental. Every scientific finding, says the physical-to-mental reductionist, is a *finding*, confirming or rebutting hypotheses and leading to theories, all of which are mental entities, which thus should be considered fundamental – not to speak of the fact that mental activity underlies any scientific activity.

So, obviously the same empirical data is compatible with two strongly differing philosophical positions. Likewise, philosophical dualists can and actually do simply claim that empirical data *per se* do not conflict with dualism (see, e.g., Lowe 2006), but rather certain presumptions and *a priori* conceptual framings *applied to* the empirical data.

It is a widely accepted cliché that Descartes' interactionist dualism was common and unquestioned at his time, and declined in plausibility as scientific progress was made. But in fact, the major systematic flaw of interactionist dualism – not being able to provide a sufficient account of mental causation – kept most of his contemporaries from endorsing Descartes' position. As Jaegwon Kim points out, "it is more than a little amazing to realize that Descartes was an exception rather than the rule, among the great Rationalists of his day, in defending mental causation as an integral element of his view of the mind" (Kim 2005, 8), in contrast to, e.g., Leibniz's parallelism and Spinoza's double aspect monism.

So, obviously, interactionist dualism, although perhaps in keeping with commonsense, has always had a hard time in the philosophy of mind – for its serious conceptual and systematical, i.e., philosophical, shortcomings –, irrespective of knowing or not knowing 20th and 21st century neuroscience.

Generally speaking, the position put forward here suggests that the ontological status of experiential properties – whether they eventually are identical or not identical with non-experiential properties, i.e., the physicalist versus the (property) dualist standpoint – will be debated on the basis of conceptual, not empirical, means.

That said, empirical data on (qualitative and quantitative) consciousness can – beyond hard philosophical categorizations – significantly shape our thinking and our conception of consciousness. From a clinical standpoint in particular striking is the idea of establishing methods to study and evaluate the neural correlates of consciousness *in a single individual*, and to determine the state or level of consciousness via neuroscientific means in cases where the individual's behavior does not give a definite indication of

their level of consciousness. E.g., proponents of the global workspace theory have argued that, on the basis of this theory, predictions on an individual's consciousness level (and on the future clinical outcome) can be made, and provided evidence for the evaluation of consciousness on the basis of event-related EEG potentials (Bekinschtein et al. 2009; Faugeras et al. 2012). Integrated information theory has also been claimed of being able to extrapolate and make inferences about consciousness in clinically or behaviorally unclear cases (Tononi & Koch 2015), although the theory provides a more abstract and mathematical approach.

Relevant cases for determining the level of consciousness via neuroscientific methods include neurological patients with severe clinical impairment of quantitative consciousness but however circumscribed brain lesions: empirical theories of consciousness, and the brain measuring methods involved in establishing it, could help to differentiate, e.g., cases of locked-in syndrome (i.e., a state of fully preserved consciousness without ability to behaviorally respond) from unresponsive wakefulness (i.e., a state without consciousness) – in particular cases of complete locked-in syndrome, where not only horizontal, but also vertical eye movements are affected due to lesions extending from the pons to the ventromedial midbrain (Bauer et al. 1979; Das et al. 2021), and thus even the minimal behavioral response usually indicating conscious experience in locked-in patients is not preserved. Other potential applications of empirical measurements of (levels of) consciousness include newborn babies, animals, and possibly one day even complicated machines and artificial intelligence (Tononi & Koch 2015).

Another fascinating way how neuroscience could change our understanding of consciousness (and of ourselves as conscious beings) is that it could equip us with methods to distinguish phenomenal consciousness itself from the mere "reportability" of conscious content. Consciousness – so it seems reasonable to assume – is about what is conscious at the very moment of experiencing it, and not about what can be reported afterwards: it could be that our conscious experience is much richer than what we can report and what we can even remember one moment later – and neuroscience will enable us to study these phenomenal qualities that are even inaccessible to the subject's own thinking (Lamme 2010).

Here, however, conceptual stipulations and philosophical presumptions are looming again: take, for the sake of the argument, an eliminativist stance towards qualia. Then, reportability of conscious content simply is *all there is* to consciousness: it makes no sense to speak of phenomenal awareness irrespective of its being remembered or reported, because such processes were indiscernible from unconscious processes. Whether to term them conscious (but instantly forgotten) or unconscious does not pick out any

real difference in the world; it's simply an arbitrary decision about naming things. This, of course, is an untenable claim to the realist about qualia: whether or not we have qualitative consciousness at a given moment – irrespective of its consequences – makes a huge, perhaps maximal, difference.

Agency, Freedom, and Empirical Determinants

The topic of free will – of free acts and free decisions – has been widely debated in philosophy, although mostly not as part of the mind-body problem, which mainly focusses on consciousness in the strict sense. (It may be, however, that the problem of free will *should* be considered as integral part of the mind-body problem and as its second "dimension" next to conscious experience; see, e.g., Griffin 1998, for putting forward this view.) The concept of *freedom* can be supplemented by the concept of *agency*: what does it mean to "act," and what distinguishes actions from mere happenings (perhaps with an intermediate step between actions and happenings called "doings"; see, e.g. Nida-Rümelin 2007)?

Much of the debate has focused on whether free will requires indeterminism (i.e., the incompatibilist view of freedom and determinism, or the libertarian view of freedom), or whether it is compatible with determinism (i.e., the compatibilist view). Sometimes it is argued that even (an adequate concept of) agency, and not only of free will, requires a form of indeterminism (Stewart 2012).

I will not opt for either side here (arguments for both sides will be presented in the following), but try to relate conceptual philosophical issues to some empirical findings and their implications on acting and (free) decision making.

Arguably, the best known and most influential neurophysiological experiments on free will and decision making are the Libet experiments conducted in the 1980s, showing, in a voluntary and "spontaneous" movement task, the scalp-measured Bereitschaftspotential – indexing (supplemental) motor cortex activation – to precede the conscious decision to move (as determined through the subject's recall of the spatial "clock-position" of a revolving spot) by several hundreds of milliseconds (Libet et al. 1982 1983; Libet 1985).

Many subsequent studies then have targeted the unconscious precursors of conscious decisions (and addressed some of the methodological problems of Libet's original experiments), generally confirming that predictors of the outcome of a decision can be detected some time before the decision becomes conscious. E.g., Fried et al.

(2011) report, in an intracranial EEG study, single supplementary motor area neurons to predict an impending decision to move several hundreds of milliseconds prior to volition. Notably, Soon et al. (2008) found the local spatial pattern of fMRI responses in frontopolar and parietal cortical areas to differ with respect to a spontaneous voluntary button press with the right or left index finger that occurred up to 10 seconds later.

According to one interpretation of the Libet and following experiments, these findings exclude a causal role of the conscious decision, making it a mere epiphenomenon that occurs only after the "true" decision has already been made unconsciously – albeit Libet himself favored a "veto" view of the (conscious) will, according to which it still holds the "possibility of stopping or vetoing the final progress of the volitional process" (Libet 1999). (The possibility that the veto decision itself would be preconditioned or determined by preceding unconscious processes was obviously dismissed by Libet.)

The crucial *conceptual* question, however, of whether an adequate definition of freedom and free decision-making should exclude the preconditioning of an act or a decision by antecedent events is hardly satisfactorily addressed in most of the empirical literature (but see, e.g., Roskies 2010, for a discussion).

Libet obviously favored an incompatibilist view of freedom and determinism, stating that his "operational definition of free will […] was in accord with common views" and that a position according to which "consciously willed acts are fully determined by natural laws […] would make free will illusory" (Libet 1999); however, he seems not to be discussing arguments for this pre-empirical choice itself.

The philosophical alternative between compatibilism – the view that free will and determinism are compatible – and libertarianism – the view that they are not; and that we must assume indeterminism to secure free will – sometimes is formulated such that our natural intuitions, or common sense, are pushing us towards libertarianism, whereas the sciences (not only neuroscience; also such other empirical disciplines as genetics or developmental psychology, cultural and historical sciences and so forth) push us towards compatibilism, pointing to the multiple antecedents and constraints to which our actions are subject.

Although there is some truth to this statement, I don't believe this is the best way to put it. It is true that strong intuitions speak in favor of incompatibilism; however, evenly strong intuitions speak in favor of compatibilism – so that the compatibilism-libertarianism alternative is best described as a struggle between *different intuitions* about ourselves as acting and deciding subjects. I will briefly elaborate the two major conceptual, i.e., pre-empirical, pro-incompatibilism and pro-compatibilism arguments.

The *Consequence argument* for incompatibilism is usually attributed to Peter van Inwagen, although it can be traced back to scholasticism and probably to ancient Greek philosophy (Jäger 2013). It focusses on the idea of alternative possibilities to act, which is core to our sense of freedom. If determinism is true – says the argument –, then each state of the world, including my actions, is necessitated by its preceding state (and by the laws of nature), and the preceding state again is necessitated by the pre-preceding state, and so forth, up to events in the remote past. If, however, my acts and decisions are an inevitable consequence of events in the remote past (and of the laws of nature), both of which are not up to me, then also my acts (and my decisions) are not up to me, and I cannot be attributed free will (Van Inwagen 1983).

The pro-compatibilist argument – sometimes referred to as *Mind argument*, as it was formulated in several publications in the namesake journal – acknowledges that the idea of alternative (originally called "alternate," see Frankfurt 1969) possibilities is crucial to our sense of freedom, but however argues that it is essentially misunderstood in the incompatibilist paradigm. Surely, the compatibilist argument goes, we have the freedom to act otherwise – but only if the past, i.e., our thoughts, our intentions and decisions, would have been otherwise. The libertarian view evoking an *indeterministic* relationship between reasoning and deciding, deciding and acting – stipulating that a given past, including my own reasoning and deciding, can lead to different, "branching" futures of acting – provides no coherent idea of freedom, but only of randomness and arbitrariness. An adequate concept of freedom, thus, it not only consistent with but *requires* determinism; and the difference between freedom and unfreedom lies in the right *kind* of determination of the present and future by the past.

So, the Consequence argument relies on the strong intuition that, given exactly the same past including our own reasoning, we have different options of deciding and acting, whereas the Mind argument relies on the equally strong intuition that the very same reasoning leading to differing decisions and actions would not constitute freedom, but rather a kind of randomness seeming undesirable and perhaps upsetting.

Another question related to determination is whether free choices must not be *predictable*. E.g., Libet (1999) seems to adopt this view, stating – after having asserted that free will is incompatible with determinism – that "even if events are not predictable in practice, they might nevertheless be in accord with natural laws and therefore determined," suggesting that predictability is a sufficient (although not necessary) condition for an event, such as a human action or decision, not to be counted as free.

The view that predictability excludes freedom might, however, be questioned: why should a predictable or foreseeable action be less free than an unpredictable, spontaneous

one? Freedom seems to be confused with arbitrariness here. After all, it's easy to imagine cases with a near-to-100% predictability of a person's action – being the prediction based on neuroscience or on common sense – with being it nonetheless a prime example for a free action (see, e.g., Nida-Rümelin 2018, for several examples of predictable and "psychologically determined" [as opposed to "microphysically determined"] free actions). It seems thus fair to say, I would like to stress, that strong intuitions are pointing us to the compatibility of freedom and predictability.

However, my point here lies not in putting forward compatibilist arguments against the incompatibilism that Libet and others tacitly seem to presuppose – I'm remaining neutral here with regard to the compatibilist vs. libertarian alternative –, but in emphasizing that the *most relevant* questions about free will and agency are of conceptual nature, i.e., in determining whether conditioning by the past renders freedom impossible, or, on the contrary, freedom precisely consists in the right kind of human actions and decisions being conditioned by preceding events.

So – if conceptual arguments and assumptions (logically, not temporarily) precede empirical data, what then, conversely, is the impact of empirical results on the conceptual debate? Universal determinism, to be sure, is a metaphysical, not a physical, thesis; however, a scientific picture of the world, as fleshed out by the numerous empirical disciplines, might still make us inclined towards compatibilism rather than libertarianism: as the empirical data – of neuroscience, to begin with, but also of many more empirical disciplines including the social sciences – emphasizes the many factors that condition human deciding and acting, the compatibilist picture, with its localization of freedom *within* an overall framework of past events necessitating present and future events, although not directly empirically provable, may look more attractive.

Conclusions

Addressing consciousness – both the ontology of conscious experience itself as well as the (free) agency of consciously experiencing subjects – in a multidisciplinary manner from both a philosophical and a scientific stance reveals that the problems or questions of both disciplines are *distinct*: philosophical concepts are typically not themselves empirically testable; and one discipline is unlikely to replace the other.

While philosophy's approach will ponder the impact of empirical data on (pre-empirical) philosophical concepts, the neuroscientific stance will elaborate the implications and consequences of empirical findings starting from a given concept.

An example for the latter, in the case of consciousness, might be Tononi's approach to the integrated information theory: instead of "trying to 'distill' mind out of matter" (which – if possible or not – would be a conceptual undertaking), he suggests to "start from consciousness itself, by identifying its essential properties, and then ask what kinds of physical mechanisms could possibly account for them" (Tononi & Koch 2015).

Maybe philosophical and empirical theories should even be combined to sufficiently account for conscious experience. The philosopher Gregg Rosenberg opts for combining his theory of natural individuals (TNI) – linking intrinsic experiential properties and causation within a complex panexperientialist hierarchy of individuals – with the integrated information theory (IIT): "By joining TNI and IIT, one gets both a metaphysics and a physics for understanding the presence of consciousness in our world: *why* is it present; *where* is it present; and *how much* is present. TNI adds to IIT a deeper but still naturalistic explanation of why integrated information is experiential." (Rosenberg 2016, 171)

Whether those alliances will prove to be fruitful is itself an empirical question (given an adequate empirical definition of "fruitful"), that the upcoming decades of research will be answering. In any case, multidisciplinary and intersectional perspectives including both philosophy and neuroscience are needed to pinpoint the concepts, correlates and causes of conscious experience.

References

Bauer, G., Gerstenbrand, F., & Rumpl, E. 1979. "Varieties of the Locked-in Syndrome." *Journal of Neurology* 221 (2): 77–91.

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., & Naccache, L. 2009. "Neural Signature of the Conscious Processing of Auditory Regularities." *Proceedings of the National Academy of Sciences* 106 (5): 1672–1677.

Block, N. 1978. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9: 261–325.

Block, N. 1995. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18 (2): 227–247.

Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford Paperbacks.

Crick, F., & Koch, C. 1998. "Consciousness and Neuroscience." *Cerebral Cortex* 8 (2): 97–107.

Das, J. M., Anosike, K., & Asuncion, R. M. D. 2021. *Locked-in Syndrome*. *StatPearls [Internet]*. StatPearls Publishing.

Dehaene, S., & Changeux, J. P. 2011. "Experimental and Theoretical Approaches to Conscious Processing." *Neuron* 70 (2): 200–227.

Dehaene, S. 2014. *Consciousness and the Brain: Deciphering How the Brain Codes our Thoughts*. New York: Penguin.

Faugeras, F., Rohaut, B., Weiss, N., Bekinschtein, T., Galanaud, D., Puybasset, L., ... & Naccache, L. 2012. "Event Related Potentials Elicited by Violations of Auditory Regularities in Patients with Impaired Consciousness." *Neuropsychologia* 50 (3): 403–418.

Frankfurt, H. 1969. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66 (23): 829–839.

Fried, I., Mukamel, R., & Kreiman, G. 2011. "Internally Generated Preactivation of Single Neurons in Human Medial Frontal Cortex Predicts Volition." *Neuron* 69 (3): 548–562.

Griffin, D. R. 1998. *Unsnarling the World-Knot: Consciousness, Freedom, and the Mind-Body Problem*. University of California Press.

Jackson, F. 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 32 (127): 127–36.

Jäger, C. 2013. "Das Konsequenzargument." In *Klassische Argumentationen der Philosophie*, 275–296. Brill Mentis.

Kim, J. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Koch, C., Massimini, M., Boly, M., & Tononi, G. 2016. "Neural Correlates of Consciousness: Progress and Problems." *Nature Reviews Neuroscience* 17 (5): 307–321.

Lamme, V. A. 2006. "Towards a True Neural Stance on Consciousness." *Trends in Cognitive Sciences* 10 (11): 494–501.

Lamme, V. A. 2010. "How Neuroscience Will Change Our View on Consciousness." *Cognitive Neuroscience* 1 (3): 204–220.

Libet, B., Wright, E. W., Jr. & Gleason, C. A. 1982. "Readiness-Potentials Preceding Unrestricted "Spontaneous" vs. Pre-Planned Voluntary Acts." *Electroencephalography and Clinical Neurophysiology* 54: 322–335.

Libet, B., Gleason, C. A., Wright, E. W. & Pearl, D. K. 1983. "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activities (Readiness-Potential); the Unconscious Initiation of a Freely Voluntary Act." *Brain* 106: 623–642.

Libet, B. 1985. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *Behavioral and Brain Sciences* 8 (4): 529–539.

Libet, B. 1999. "Do We Have Free Will?" *Journal of Consciousness Studies* 6 (8-9): 47–57.

Lowe, E. J. 2006. "Non-Cartesian Substance Dualism and the Problem of Mental Causation." *Erkenntnis* 65 (1): 5–23.

Mashour, G. A., Roelfsema, , Changeux, J. P., & Dehaene, S. 2020. "Conscious Processing and the Global Neuronal Workspace Hypothesis." *Neuron* 105 (5): 776–798.

Nagel, T. 1974. "What is it like to be a bat." *Readings in philosophy of psychology* 1: 159-168.

Nida-Rümelin, M. 2007. "Doings and Subject Causation." *Erkenntnis* 67 (2): 255–272.

Nida-Rümelin, M. 2018. "Freedom and the Phenomenology of Agency." *Erkenntnis* 83 (1): 61–87.

Rosenberg, G. 2016. "Land Ho? We Are Close to a Synoptic Understanding of Consciousness." In *Panpsychism: Contemporary Perspectives*, 153–78. Oxford: Oxford University Press.

Roskies, A. L. 2010. "How Does Neuroscience Affect Our Conception of Volition?" *Annual Review of Neuroscience* 33: 109–130.

Seth, A. K., & Bayne, T. 2022. "Theories of Consciousness." *Nature Reviews Neuroscience* 23: 439–452.

Siewert, C. 1998. *The Significance of Consciousness*. Princeton: Princeton University Press.

Soon, C. S., Brass, M., Heinze, H. J., & Haynes, J. D. 2008. "Unconscious Determinants of Free Decisions in the Human Brain." *Nature Neuroscience* 11 (5): 543–545.

Steward, H. 2012. *A Metaphysics for Freedom*. Oxford: Oxford University Press.

Tononi, G. 2008. Consciousness as Integrated Information: A Provisional Manifesto. *The Biological Bulletin* 215 (3): 216–242.

Tononi, G., & Koch, C. 2015. "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society B: Biological Sciences* 370 (1668): 20140167.

Tononi, G., Boly, M., Massimini, M., & Koch, C. 2016. "Integrated Information Theory: From Consciousness to Its Physical Substrate." *Nature Reviews Neuroscience* 17 (7): 450–461.

Van Inwagen, P. 1983. *An Essay on Free Will.* Oxford: Oxford University Press.

cognethic.org