Fragments

McGill Undergraduate Journal of Philosophy

Issue 35
2021

Team

Editors-in-Chief

Matthew Gery

Alex Louis

Editors

Castilleja DeMarco

Lintao Dong

John Jakob Etter

Dante Fasulo

Emma Slack-Jørgensen

Cover Artist

Sarah Gery

Contents

An Anarchist's Challenge to the Legal Authority of States

Finn Boyle

One argument which appears in the diverse field of anarchist thought holds that laws which stem from coercive institutions like the state are fundamentally illegitimate, amounting to nothing less than crude assertions of power that effectively constrain personal autonomy and freedom. This argument suggests that since laws are produced by the state, an inherently coercive institution, and not ratified by the consent and consensus of the governed, their authority is illegitimate. It would then logically follow that there is no legitimate justification for following the rule of law of any state merely on the basis that it is the law. However, while this implies that a law should not be obeyed due to its coercive nature, a coercive law could indeed have legitimate reasons to be obeyed. Furthermore, the valorization of individual autonomy and freedom at the expense of any form of coercion ignores the moral utility said coercion often serves, merely arguing that because something is 'coercive' it is therefore illegitimate. While criticisms can be made about the legitimacy of law, its coercive nature is not what determines said legitimacy, as the potential moral outcomes of coercion can justify its use.

Our anarchist's[1] conception of law aligns very closely with the defini-

[1] To be clear: whenever I refer to an "anarchist" or "our anarchist" I am referring specifically to the hypothetical view put forward in the introduction, rather than any particular anarchist thinker or

tion given by John Austin. For them, a law is any "rule laid down for the guidance of an intelligent being by an intelligent being having power over [them]." In short, for both Austin and our anarchist, law is little more than a "crude assertion of power."[2] This thesis is bolstered by the arguments of the American Legal Realists, who collectively held that law is both "causally" and "rationally indeterminate"—causally indeterminate in that "legal reasons did not suffice to explain" why law is applied in the ways it is, and rationally indeterminate "in the sense that the available class of legal reasons did not justify a unique" outcome.[3] If law is a mere 'crude assertion of power' by one being unto another, then it would logically entail a certain sense of indeterminacy, as the law would stem from the indeterminable judgment of a single or multiple so-called "intelligent [beings]."[4]

One of the most common counter-arguments to this view of the law is that this definition makes no distinction between any form of violently enforced authority. Legal theorist H.L.A Hart criticized this conception of law as a "gunman situation… writ large." He gives the example: "A orders B to hand over his money and threatens to shoot him if he does not comply. According to the theory of coercive orders this situation illustrates the notion of obligation or duty in general."[5] This was meant as a refutation of Austin's command theory of law, as obviously a gunman ordering his victim to hand over his money under threat of violence isn't exercising legal authority. However, our anarchist might fully agree with this analysis, arguing that there indeed is no meaningful difference between a gunman robbing someone with an implicit threat of violence and the law threatening sanctions against a person. They could argue that this is precisely why the legal authority of states is illegitimate, as it is still un-

school of thought.

[2] John Austin, "Lecture I," in *The Province of Jurisprudence Determined* (Cambridge: Harvard University Press, 1998).

[3] Brian Leiter, "American Legal Realism," *University of Texas Public Law Research Paper*, no. 42 (October 2002).

[4] John Austin, "Lecture I."

[5] H.L.A. Hart, *The Concept of Law* (Oxford: Oxford University Press, 1994).

deniably coercive and cannot be meaningfully distinguished from any other crude assertion of power. Therefore, if one were to refute our anarchist's argument, they would have to find some way to differentiate legal coercion from other forms of coercion.

Hart argues that the difference between the "gunman situation" and the application of real law is the presence of obligation. He argues that situations in which one's coercive authority is enforced can result in someone being "obliged" to obey, whereas the authority of law institutes an "obligation." The difference between the two is one of universality, with legal obligations being "habitually obeyed and…general," as they prescribe "courses of conduct" and "not single actions."[6] Whereas a person being mugged by a gunman might be obliged to comply out of force, this act of obliging is particular and has no *universal* sanction attached to it. For Hart, this is the key difference between a random 'crude assertion of power' and the law—whereas assertions of power might be particular and *ad hoc*, law is a prescribed set of sanctions applied to a large group of people on a systematic basis.

However, a hypothetical anarchist might counter Hart by citing the work of the American Legal Realists. As previously stated, the American Legal Realists held that law was functionally "indeterminate" due to the large host of non-legal factors involved in various legal processes. While law theoretically may be fair and equal, in practice it is often contradictory and unequal in application. American Legal Realist Brian Leiter cites Judge Chancellor Kent, who claimed to approach legal cases by looking for "where justice lay…. I then sat down to search the authorities…. *I almost always found principles suited to my view of the case*."[7] Our anarchist would argue that, since law is not enforced fairly in practice, it is therefore indeterminate and illegitimate and remains little more than a crude assertion of power.

However, the extent to which the law is truly indeterminate has been heavily debated. Ronald Dworkin claimed that propositions of law are "interpre-

[6] Ibid.

[7] Brian Leiter, "American Legal Realism." Original emphasis.

tive of [their own] legal history" and therefore not entirely indeterminate. A proposed law cannot completely contradict an existing law, since the resulting situation would be unsustainable and legal institutions would have to resolve the contradiction. No law comes into being *ex nihilo*; its range of possible content is bound by its own history and predecessors, and therefore cannot be entirely indeterminate.[8]

An example one might cite to further argue against the indeterminacy of law as outlined by the American Legal Realists is a 1949 West German trial of a woman for "illegally depriving a person of his freedom." She had reported her husband to the Nazis in 1944 for speaking ill of the Third Reich, a crime for which he was sentenced to execution (although his sentence was later commuted). While she believed she had acted in accordance with the law at the time, she was found guilty because her actions ran contrary to the "German Criminal Code of 1871 which had remained in force continuously since its enactment." According to Hart, two laws contradicted each other: one banned criticism of the Nazi regime and the other prohibited the deprivation of a person's freedom. Rather than letting this contradiction fester, the West German state resolved it on the professed basis of "the sound conscience and sense of justice of all decent human beings." [9]

This citation of morality presents another challenge to our anarchist's assertion that law is illegitimate. How can a law backed by the "sense of justice of all decent human beings" be illegitimate? If law isn't wholly indeterminate and could be backed by near-universal moral intuitions, then surely we could distinguish such morally legitimate laws from the aforementioned 'crude assertions of power.' If a law corresponds with a universal sense of morality, it would be folly to label it 'illegitimate.' Accordingly, we can legitimate the legal authority of states through appeal to a more fundamental moral legitimacy.

While following a law which happens to be moral because it is moral does not

[8] Ronald Dworkin, "Law as Interpretation," in *Texas Law Review* 60, no. 3 (March 1982).

[9] H.L.A. Hart, "Positivism and the Separation of Morals," *Harvard Law Review* 71, no. 4 (February 1958): 593-629.

entail 'following the law' *per se*, as what is being obeyed is morality itself rather than the law, it does confer a certain sense of legitimacy. Rev. Dr. Martin Luther King, Jr. differentiated between "two types of laws: just and unjust…. A just law is a man made [sic] code that squares with the moral law…. An unjust law is a code that is out of harmony with the moral law." Claiming that he "would be the first to advocate obeying" *just* laws, King argued that one could be found in the integration policy that resulted from the Supreme Court's 1954 *Brown v. Board of Education of Topeka* decision.[10] The Supreme Court's desegregation order is by no means uncoercive—forcing public schools to desegregate is an inherently coercive act—but that coercion does not make the law immoral or illegitimate. Indeed, in this case coercion served an undoubtedly moral purpose: combatting racism. If a coercive law can serve a purely moral purpose and square itself with the moral law, then at the very least its coercive enforcement would be morally defensible. King's concept of the just law provides us with a second moral basis upon which we can secure the legitimacy of state legal authority.

The core issue at play in our anarchist's challenge to legal authority is the implicit assumption that coercion and constraining individual freedom and autonomy render state-backed law illegitimate. However, coercion can often serve a moral purpose, and the moral outcomes of this coercion can serve to justify its application. As previously stated by Hart, a defining feature of law is its ability to impose obligations on a group of people. An example of such an obligation might be a workplace safety law, wherein an employer who does not provide a safe work environment for their employees may be found criminally negligent and sanctioned. Such a law would effectively constrain the individual freedom of the employer to dictate their workplace as they see fit, but it would also lead to a much safer work environment and benefit more people than it would constrain. Indeed, one could make the argument that this act of coercion would overall grant more people *more* freedom, as the outcome of the law's enforcement would lead to fewer workplace accidents, which themselves could con-

[10] Martin Luther King, *Why We Can't Wait* (New York: Signet, 1964).

strain the bodily autonomy of workers by physically disabling them. Coercion, in this sense, can serve a legitimate good, and provide a social net benefit in the view of most.

However, let's assume that our anarchist finds this argument unconvincing. They could cite Joseph Raz's Paradox of the Just Government, arguing that "the moral obligations on which the claim that the law is just is founded are prior to and independent of the moral obligation to obey the law." The anarchist might claim that what is legitimate is not the act of coercion itself but what the coercion is trying to achieve. If one should respect the law due to its moral nature, rather than it simply being a law, "the obligation to obey the law is at best redundant." Yet, while the reasons one uses to obey the law could be criticized and deemed morally insufficient, the outcome of doing so would still be morally beneficial to society as a whole. Raz himself asks rhetorically, "can there be a moral obligation to perform an action if to take the existence of the obligation as one's reason for the action it enjoins would be wrong, or ill-fitting? So much for the apparent paradox of the just law."[11]

Raz's argument states that the mere existence of obligation does not negate the legitimacy of a coercive act, provided that the act is morally or justly guided. One could just as easily claim that the social enforcement of morality is coercive. Yet, the coercive nature of a socially enforced morality (e.g. society shunning thieves) does not render it illegitimate, and the same could be said of the laws of states. Our anarchist's view of coercion itself may be malformed, as they merely assume that the presence of any coercion automatically detracts from freedom. The fact that the reality of coercion is far more nuanced undermines their argument.

Furthermore, the argument that the coercive nature of the state renders its laws illegitimate also implies a paradox. If the laws of states are illegitimate because they are coercive 'crude assertions of power,' then what morally justified action should be taken to rectify this? Doubtless our anarchist would

[11] Joseph Raz, "The Obligation to Obey: Revision and Tradition," *Notre Dame Journal of Law, Ethics, and Public Policy* 1, no. 1 (1985).

call for the abolition of the state and its illegitimate law. The means by which this hypothetical state and law are abolished are not wholly important here. One could theorize a violent overthrow or a peaceful takeover of power leading to the gradual dissolution of the state. The outcome of both scenarios remains the same: a previously existing state has ceased to be. In every conceivable scenario, to abolish or attack the state in any form whatsoever requires coercion, even if said abolition or attack occurred without the use of direct physical violence.

This argument does not imply that only coercion used to remove greater coercion is justified, merely that coercion in itself is not necessarily illegitimate. If coercion is the mere forcing or pressuring of someone to perform a certain act against their will, then any act against a so-called illegitimate state would entail coercion. The need to use coercion in specific circumstances to remove greater coercion signals that coercion itself is not enough to render states—or, therefore, their laws—illegitimate. The core issue with our anarchist's challenge to legal authority is that it is based on a rejection of any form of coercion as illegitimate, when, as previously established, coercion is often necessary to achieve morally justified goals.

This is not to say that the law is an inherently good or universally legitimate institution, merely that coercion and assertions of power alone cannot determine legitimacy. State-backed law is evidently a coercive institution, but so is socially-enforced morality. Laws of the state by their nature infringe on select freedoms; yet, some freedoms contradict other freedoms, and so it is necessary that *some* will end up infringed upon (e.g. the freedom of personal autonomy of a business owner clashes with the freedom of workers from workplace accidents). Furthermore, the existence of laws backed by near-universal moral intuitions throws a spanner into the works of the law's supposed illegitimacy, as these laws cannot and should not be considered illegitimate, regardless of state enforcement.

Bibliography

Austin, John. "Lecture I." In *The Province of Jurisprudence Determined*. Cambridge: Harvard University Press, 1998.

Dworkin, Ronald. "Law as Interpretation." *Texas Law Review* 60, no. 3 (March 1982).

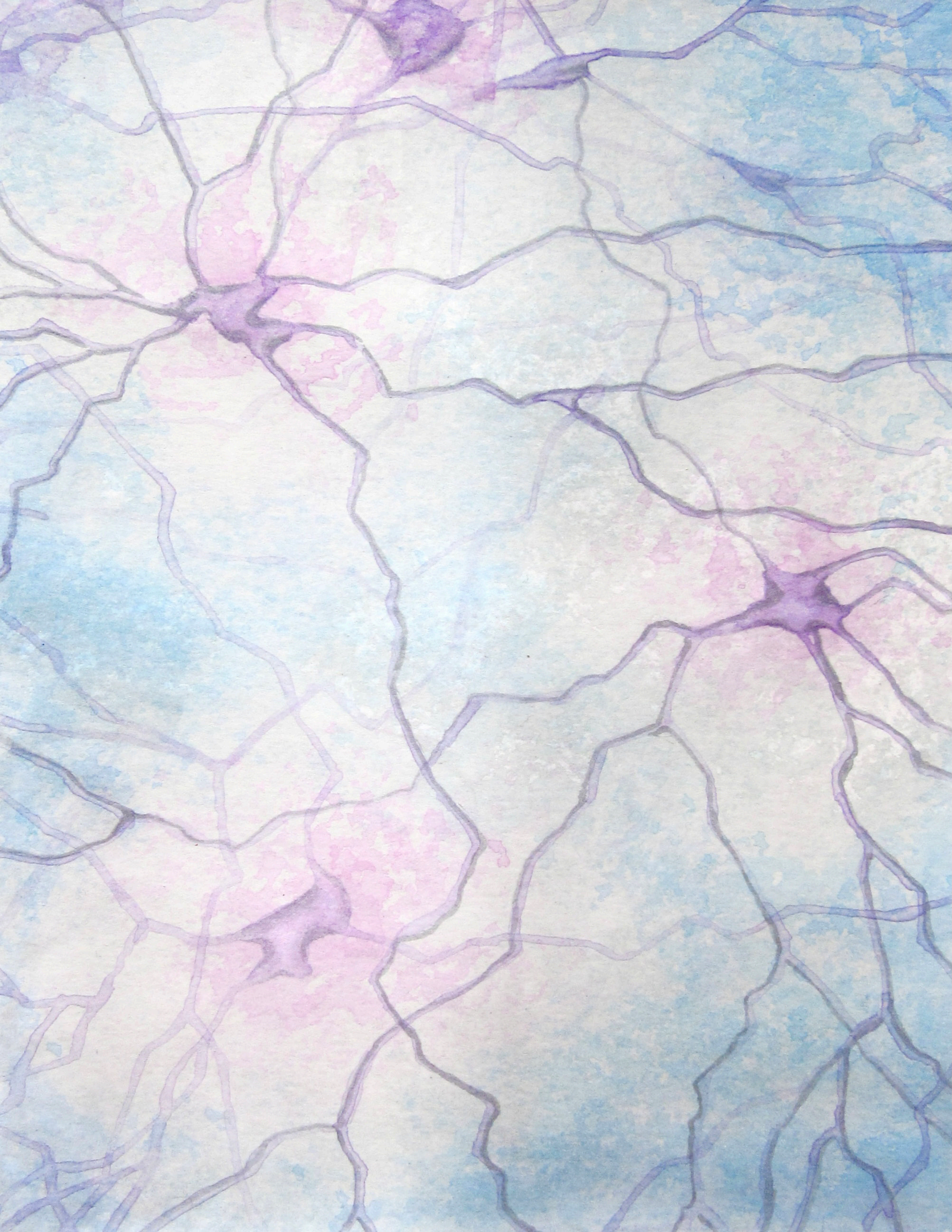
Hart, H.L.A. *The Concept of Law.* Oxford: Oxford University Press, 1994.

—. "Positivism and the Separation of Morals." *Harvard Law Review* 71, no. 4 (February 1958): 593-629.

King Jr., Martin Luther. *Why We Can't Wait*. New York: Signet, 1964.

Leiter, Brian. "American Legal Realism." *University of Texas Public Law Research Paper*, no. 42 (October 2002).

Raz, Joseph. "The Obligation to Obey: Revision and Tradition." *Notre Dame Journal of Law, Ethics, and Public Policy* 1, no. 1 (1985).



A Means to an End: The Use of Brain Organoids in Research

Samir P. Gouin

I. *Introduction.*

Brain organoids, also referred to as mini-brains, currently resemble pale pink miniature egg yolks. They lack sulci and gyri and are quite small in size, not even approaching the intricate complexity of the brains of commonly used experimental animal models such as rats. However, many researchers are already debating the ethical repercussions of further developing brain organoids.[1] [2] [3] [4] [5] [6] To be able to use brain organoids in research to their full potential, they should resemble human brains. Brain organoids may exhibit not only structural similarities to human brains

[1] Isaac H. Chen et al, "Transplantation of Human Brain Organoids: Revisiting the Science and Ethics of Brain Chimeras," *Cell Stem Cell* 25, no. 4 (October 2019): 462-472, doi:10.1016/j.stem.2019.09.002.

[2] Insoo Hyun, J.C. Scharf-Deering and Jeantine E. Lunshof. "Ethical issues related to brain organoid research." *Brain Research* 1732 (April 2020): #146653, doi:https://doi.org/10.1016/j.brainres.2020.146653.

[3] Koplin & Savulescu, 2019 Julian J. Koplin and Julian Savulescu, "Moral Limits of Brain Organoid Research," *Journal of Law, Medicine, and Ethics* 47, no. 4 (December 2019): 760-767, doi:10.1177/1073110519897789.

[4] Andrea Lavazza, "Human cerebral organoids and consciousness: a double-edged sword," *Monash Bioethics Review* 38, no. 2 (September 2020): 105-128, doi:10.1007/s40592-020-00116-y.

[5] Andrea Lavazza and Marcello Massimini, "Cerebral organoids: ethical issues and consciousness assessment," *Journal of Medical Ethics* 44 (September 2018): 606-610. doi:10.1136/medethics-2017-104555.

[6] Megan Munsie, Insoo Hyun, and Jeremy Sugarman, "Ethical issues in human organoid and gastruloid research," *Development* 144 (March 2017): 942-945. doi:10.1242/dev.140111.

but also functional ones, such as a sense of consciousness (defined here as subjective awareness).

If accurate tests to measure attributes of consciousness were developed, which would represent a giant leap from current research, and brain organoids showed signs of consciousness, the perception of brain organoids could suddenly shift from them being inanimate pieces of tissue to being biological entities that may be capable of thought. To best address this area of ethical contention, I first establish a picture of current research and then examine the ethics of brain organoids. The ethical question my paper seeks to address is whether it is permissible to use advanced brain organoids in research. I will employ two distinct ethical frameworks to explore this question from two perspectives. Through a utilitarian perspective, the benefits, such as better development of current and new treatment methods, and the reduced necessity of animal testing outweigh the possible suffering induced in a brain organoid. Through a Kantian perspective, brain organoids do not necessitate the same moral consideration as other conscious beings due to differences in development. As part of this inquiry, I explore under what conditions brain organoids could be considered conscious. While brain organoids may develop sentience and sapience, and therefore satisfy some necessary conditions for moral consideration, their heteronomous nature nevertheless disqualifies them from it. Therefore, on the grounds of evaluation through the utilitarian and Kantian frameworks, I propose that the use of brain organoids in research is ethically permissible.

II. *Emergence of Brain Organoids in Research.*

Seemingly out of science fiction, organoids are being grown in many labs around the world to mimic the function of natural human organs. By using stem cells, researchers have been able to synthesize specific organs ranging from kidneys to stereocilia in the inner ear. Many are hopeful that these lab-grown organs will replace many currently used research models. What makes organoids special is their ability to self-differentiate in a fashion similar to the development of normal human tissues. Like seeds to a tree, researchers could plant stem cell clusters and allow them to grow into complex structures, thus facilitating organoid harvest for research.

In 2005, Yoshiki Sasai and his team were the first group successful in synthesizing 3D neural tissue from rodent stem cells.[7] In 2008, the same group used human embryonic stem cells to grow cerebral neural tissue.[8] These foundational studies paved the way for advancements in supporting neurogenesis and corticogenesis, specifically by improving suspension cultures. Since then, researchers have been able to create neural areas such as parts of the thalamus, cerebral cortex, and hippocampus, as well as brain organoids.[9] [10] [11]

One of the main challenges ahead is to devise systems akin to the roles of our ventricular and circulatory systems that can nourish the growth of brain organoids. Without proper supporting systems, brain organoids are arrested in their capacity to develop further. When this challenge is circumvented, brain organoids can be expected to develop into sophisticated structures. The next challenge would be to compare the neural activity of brain organoids to that of human brains. Current techniques such as electrode recording, calcium imaging, fluorescent tagging, and gene studies could help researchers map the functionality of a brain organoid.

If functional similarities to human brains are discovered in brain organoids,

[7] Watanabe et al., "Directed differentiation of telencephalic precursors from embryonic stem cells," *Nature Neuroscience* 8, no. 3 (February 2005): 288-296, doi:10.1038/nn1402.

[8] Eiraku et al., "Self-Organized Formation of Polarized Cortical Tissues from ESCs and Its Active Manipulation by Extrinsic Signals," *Cell Stem Cell* 3, no. 5 (November 2008): 519-532, doi:https://doi.org/10.1016/j.stem.2008.09.002.

[9] Atsushi Shiraishi, Keiko Muguruma, and Yoshiki Sasai, "Generation of thalamic neurons from mouse embryonic stem cells," *Development* 144 (April 2017): 1211-1220, doi:10.1242/dev.144071.

[10] Muguruma et al., "Self-Organization of Polarized Cerebellar Tissue in 3D Culture of Human Pluripotent Stem Cells," *Cell Reports* 10, no. 4 (February 2015): 537-550, doi:https://doi.org/10.1016/j.celrep.2014.12.051.

[11] Sakaguchi et al., "Generation of functional hippocampal neurons from self-organizing human embryonic stem cell-derived dorsomedial telencephalic tissue," *Nature Communications* 6, no. 1 (November 2015): article #8896.

it would be imperative to assess whether conscious-like activities are present. This would provide a better picture of the brain organoid's sophistication and level of further similarity to a human's brain. As research studies are uncovering many more states of consciousness than previously thought, determining a precise state of consciousness of brain organoids may pose a challenge. The recent findings that patients with neurologically locked-in conditions resembling coma are still able to communicate illustrates that consciousness is not binary but a spectrum of different levels.[12] Thus, the issue of testing consciousness alone raises a slew of concerns, both scientific and ethical (and philosophical).

Currently, there are several different ways to assess attributes of consciousness including fMRIs, basic reflex tests and the interrogation of the subject. Since brain organoids would be without their own sensory systems, many routine tests will not be applicable. One possible method is the Perturbation Complexity Index (PCI). This index measures the electrical response produced by perturbing the cerebral cortex with transcranial magnetic stimulation (TMS) and is used to quantify communication among neural structures. This approach is independent of sensory systems and has been used to assess the presence of consciousness characteristics in unresponsive subjects.[13] These techniques, increasingly refined to measure the consciousness of brain-injured humans, will help set the stage to evaluate the consciousness of brain organoids and allow a more informed discussion of their use in research. However, these advances in measuring attributes of consciousness will require adaptation to the brain organoid model.

III. *Sentience and Sapience of Brain Organoids.*

The concept of moral status is used throughout the ethical discussion of brain organoids. Due to its ubiquity, I would like to define it as an organism's prop-

[12] Adrian Owen, *Into the Gray Zone: A Neuroscientist Explores the Border Between Life and Death* (New York: Scribner, 2017).

[13] Lavazza, "Human Cerebral Organoids."

erty of meriting moral consideration (also known as the consideration of a being's welfare). For the purposes of this paper, moral status is considered as a graded scale based on the sophistication of organisms' mental faculties, such that an adult human has a higher moral status than a pigeon.[14] These comparisons are often made between species (i.e., so that a comatose or disabled human has the same status as fully functioning human). The level of moral status is influenced by an organism's sentience, that is, its capacity to confer value and perceive. This can be further defined as the ability to hold values and interests (practical reason) and have morally relevant features like cognitive ability (sapience).[15] There are various other criteria that philosophers have used to justify bestowing moral status, such as the sense of justice and a conception of good, but sentience, practical reason, and sapience are the most consistent and relevant across brain organoid research.[16]

As it is an indication of sapience and sentience, the predominance of a consciousness type traditionally bears significant influence on an organism's moral status. The definition of consciousness differs but it is often centered around the concept of awareness. Awareness is critical for responsiveness, as this has a direct influence on an organism's capacity for decision-making and self-government. Thus, consciousness, in some capacity, is necessary but not sufficient for autonomy. While the methods to test consciousness are incomplete, they could eventually be used to attribute moral status.

IV. *Discussion: Ethical Frameworks.*

There are significant challenges and areas of uncertainty to overcome to

[14] Jeff McMahan, *The Ethics of Killing: Problems at the Margins of Life* (New York: Oxford University Press, 2002). This is comparable to McMahan's hierarchy of being. I.e., based on a series of evaluations such as assessing the level of well-being, intrinsic potential, and psychological capacity, it is possible to rank the moral status of beings.

[15] Nick Bostrum and Eliezer Yudkowsky, "The ethics of artificial intelligence," in *The Cambridge Handbook of Artificial Intelligence,* ed. Keith Frankish and William M. Ramsey, 316-334 (Cambridge: Cambridge University Press, 2014).

[16] John Rawls, *Political Liberalism* (New York: Columbia University Press, 1993).

develop a sophisticated brain organoid. Regardless, due to the current primacy of the field coupled with a high rate of growth, it is probable that brain organoids will improve. To explore the ethics and argue the permissibility of using brain organoids in research, I apply two frameworks: utilitarianism and Kantian deontology. By reaching the same conclusion through distinct ethical frameworks, I argue that brain organoids should be used in research.

(A) *Utilitarian Framework.*

Utilitarianism is centered around derived utility. That is, an action is defined as good or bad depending on whether it maximizes the greater good for the most people possible.[17]

I will apply this framework to determine whether the infliction of suffering on a brain organoid is worth the negative consequences. It is highly likely based on current research that there will be many scientific benefits of brain organoid research. As studies have shown, many researchers are overcoming the hurdles associated with brain organoid research and are producing results based on rudimentary non-conscious models. Discrete neural tissues have already been used to investigate diseases and disorders. For example, the effect of the Zika virus on neural development has been studied in greater depth. Researchers were able to test drugs to combat Zika virus microcephaly by identifying the point of entry into neural stem cells that the virus exploits.[18] Other treatments have been explored regarding Alzheimer's disease and amyotrophic lateral sclerosis.[19] In regard to these disorders and others, brain organoids provide a means to study human-specific pathology beyond the current

[17] Editors of Encyclopaedia Britannica, "Consequentialism," *Encyclopædia Britannica* (2009), https://www.britannica.com/topic/consequentialism.

[18] Tomasz J. Nowakowski et al., "Expression Analysis Highlights AXL as a Candidate Zika Virus Entry Receptor in Neural Stem Cells," *Cell Stem Cell* 18, no. 5 (May 2016): 591-596, doi:10.1016/j.stem.2016.03.012.

[19] Tsutomu Sawai et al., "The Ethics of Cerebral Organoid Research: Being Conscious of Consciousness," *Cell Stem Cell* 13, no. 3 (September 2019): 440-447, doi:10.1016/j.stemcr.2019.08.003.

capabilities of 2D models and animals. This will help provide insight into methods to combat many diseases and disorders.

While benefits can be speculated, conscious brain organoids with functioning neural circuits could help us better understand the mind-body connection, a preliminary step to more effectively treating psychological disorders, testing new medications and much more. These benefits each have the capacity to aid many, such as patients with neuropsychological disorders, at the cost of potentially harming one being. From a utilitarian perspective, the harm inflicted on a few, the brain organoids, is outweighed by the greater good of many.

The use of brain organoids may conflict with the interests and desires of brain organoids resulting from its self-awareness. Researchers have suggested that consent may be given through the stem cell donor. However, this indirect form of consent cannot accurately convey the brain organoid's interests, in the same way a parent's consent does not always reflect the best interests of their child. Hence, in order to subject a conscious brain organoid to research, we must consider whether its moral status is significant and whether we have a corresponding moral duty towards it.

As well, while many could be aided by brain organoids, we must also ask if it is more ethical to use other research models instead of brain organoids. Today, animals are often used in research as the best approximation of human physiological systems. If brain organoids develop advanced neural circuits, they may replace animal models in certain areas of research (e.g. the study of human-specific neurodevelopmental disorders). From a utilitarian perspective, this could limit the unethical use of animals in research, thus achieving a greater good by virtue of using a lab-grown organism. However, the improvement of other methods, such as virtual simulations, artificial intelligence models, psychological assessments, and imaging/scanning techniques, might be able to provide on their own many of the benefits of using brain organoids. Therefore, to permit their use in research, they must have a moral status equivalent to or lower than currently used animal models.[20]

(B) *Kantian Framework.*

Clearly, the utilitarian calculus of the costs and benefits of using brain organoids does not reveal the entire picture. To address the counterpoints regarding the significance of a brain organoid's moral status, we can apply the deontological Kantian framework. This framework allows a closer look at using the means (using brain organoids) rather than placing emphasis on the ends (such as the large scale-outcomes of their use). Kantian deontology revolves around the concept of moral duty based on categorical imperatives—constant universal rules. Moral duty entails the consideration of the needs and desires of others beyond treating them like means to an end. For example, our duty to respect our peers stems from obligations to rational beings as a whole and not the intrinsic value of individual peers. Moral duty is not unidirectional. Instead, it depends on the reciprocation of others who are equally subject to these obligations. This concept extends beyond Kantian deontology and the capacity to reciprocate often serves as a basis to attribute moral responsibilities. Non-rational beings such as plants and animals cannot uphold duties to humans and thus, do not have duties associated with themselves. Underlying this dynamic is the concept of free will. Kant defined this as "a kind of causality that living beings exert if they are rational, and when the will can be effective independent of outside causes acting on it, that would involve this causality's property of freedom."[21] Ultimately, autonomy, or self-government, is critical to enabling beings the ability to oblige and uphold maxims.

As part of the Kantian framework, it is important to examine the moral status of brain organoids. This can be approximated through an examination of their

[20] While there is no consensus regarding the consciousness of animals, this categorization is based on cognitive abilities that are thought to reflect consciousness such as goal setting/recognition tests used to measure self-consciousness.

[21] Immanuel Kant, *Groundwork for the Metaphysics of Morals,* trans. Jonathan Bennett (unpublished manuscript, 2017), typescript, 41, https://www.earlymoderntexts.com/assets/pdfs/kant1785.pdf.

consciousness and how it may arise. Brain organoids without neural activity should warrant a moral status similar to other tissues originating from humans. This is how many researchers currently treat brain organoids.

In developing human brains, sensory stimulus allows the establishment of networks that give rise to cognitive functions.[22] [23] [24] This sensory stimulation is needed to feel pleasure and pain and develop self-consciousness through awareness. As brain organoids do not have sensory systems, artificial inputs may be required to form the sophisticated neural activity related to cognition. Further studies are needed to determine whether the maturation of brain organoids plateaus in the absence of sensory stimulation.

To mimic the natural mechanisms of sensation, researchers could use an advanced form of electrical stimulation. Similar techniques have already been used in a variety of technologies, including cochlear and visual cortical implants used to provide sensory information in the absence of functioning sensory organs. The concept of artificial neural stimulation extends beyond modern-day brain organoid research. Philosopher Hilary Putnam proposed the 'brain in a vat' thought experiment in 1981. Like René Descartes' conception of a demon fabricating our world, this experiment was conducted to explore skepticism of the empirical world. In this thought experiment, a fully functioning brain is placed in a vat. The brain's nerves are attached to a computer and stimulated with electrical pulses, producing a complete sensory perception of reality for the brain.[25] While this scenario is far-fetched, it illustrates how consciousness could be artificially simulated.

The need for brain organoids to receive artificial stimulation in order to establish neural networks reflects the organoids' total passivity, and consequently their

[22] Lavazza, "Human Cerebral Organoids."

[23] Che et al., "Layer I Interneurons Sharpen Sensory Maps during Neonatal Development," *Neuron* 99, no. 1 (July 2018): 98-116.

[24] Hugo Lagercrantz and Jean-Pierre Changeux, "The Emergence of Human Consciousness: From Fetal to Neonatal Life," *Pediatric Research* 65, no. 3 (March 2009): 255-260. doi:10.1203/PDR.0b013e3181973b0d.

[25] Hilary Putnam, *Reason, Truth, and History* (Cambridge: Cambridge University Press, 1981).

heteronomy. Unlike embryos that grow and develop active responses, the capacity of brain organoids to experience phenomena is limited to the period of stimulation. Without this input, brain organoids could lose consciousness, similar to the loss of sensory input associated with comatose states. This enduring passivity never allows for any degree of independence and autonomy, and thus disqualifies brain organoids from having free will in Kant's sense. Kant explained that for morality to be upheld as a maxim for rational beings, they must have freedom.[26] However, brain organoids are not equipped with freedom and the ability to act independently. Consequently, they cannot uphold and be held to maxims.

Moral duty and moral status work in tandem. Moral status is a necessary condition for having and owing moral duties (i.e., an entity not having moral duties entails our not having moral duties towards that entity). Likewise, having and owing moral duties can indicate moral status. Contrary to this norm, brain organoids have raised the possibility that while moral duty is often intertwined with moral status, they do not necessarily imply each other. Conscious brain organoids are at the crux of this disjoint. From a neuroscience perspective, they have some degree of moral status due to resemblance to humans, but do not have moral duties associated with them. This is radically different from most beings, such as animals, who have a lesser moral status and lesser corresponding moral duties. Animals are autonomous and have established rights from which our moral duties to them flow. This balance between moral status and moral duty also extends to humans, who, unless autonomy is severely compromised, have both moral duties and moral status. In cases with severely cognitively impaired people, a reduction in moral duty can be used to reflect limited autonomy while moral status is held constant within a species.

From a strictly Kantian perspective, brain organoids would not even qualify for a moral status as autonomy is a precursor to all things of value: "Autonomy is thus the basis for the dignity of human nature [moral status] and of every ratio-

[26] Kant, *Groundwork,* 42.

nal nature."[27] In contrast to humans lacking autonomy, such as bed-bound patients, brain organoids cannot gain or lose autonomy. Therefore, brain organoids should be treated as means in themselves, rather than as ends (the way we treat peers) as they do not possess moral status.

V. ***Conclusion.***

By leveraging two distinct ethical frameworks to examine the use of brain organoids, I have reached similar moral conclusions. Due to the lack of a moral duty of brain organoids resulting from the absence of autonomy, and the potential benefits derived from their use, it is acceptable to use them in research. But beyond Kantian deontology, whether the weight of autonomy is placed in the definition of moral status or moral duty, or both, autonomy is fundamental to the moral consideration of a being. The absence of brain organoids' autonomy refutes the attribution of moral value to them that derives from their neural resemblance to humans.

Notably, this conclusion is mitigated by the indirect effects that the use of brain organoids in research may cause such as a greater ease of biohacking, transplantation, and consciousness transfer. As Kant remarked, despite a different moral status associated with animals, there can be moral restrictions on how we treat them as this would affect how we treat each other.[28] A similar line of logic could be applied to brain organoids. While it is morally permissible to use them in research, using them mindlessly may set a problematic precedent for how we handle other situations. While these issues remain separate areas of ethical contention, they are related to the implications of this paper and merit closer examination. Nonetheless, based on current research trajectories, our understanding of consciousness, and the direct effects of brain organoids, they are a valuable means to progress in neuroscience research.

[27] Ibid., 34.

[28] Lori Greun, "The Moral Status of Animals," in *Stanford Encyclopedia of Philosophy,* ed. Edward N. Zalta, last modified Fall 2017, https://plato.stanford.edu/archives/fall2017/entries/moral-animal.

Bibliography

Bostrum, Nick and Eliezer Yudkowsky. "The ethics of artificial intelligence." In *The Cambridge Handbook of Artificial Intelligence,* edited by Keith Frankish and William M. Ramsey, 316-334. Cambridge: Cambridge University Press, 2014.

Britannica, Editors of Encyclopædia. "Consequentialism." *Encyclopædia Britannica* (2009). https://www.britannica.com/topic/consequentialism.

Brownell, Celia A., Stephanie Zerwas, and Geetha B. Ramani. "So big: the development of body self-awareness in toddlers." *Child Development* 78, no. 5 (September/October 2007): 1426-1440. doi:10.1111/j.1467-8624.2007.01075.x.

Che, Alicia, Rachel Babij, Andrew F. Iannone, Robert N. Fetcho, Monica Ferrer, Conor Liston, Gord Fishell, and Natalia V. De Marco Garcia. "Layer I Interneurons Sharpen Sensory Maps during Neonatal Development." *Neuron* 99, no. 1 (July 2018): 98-116. doi:10.1016/j.neuron.2018.06.002.

Chen, H. Isaac, John A. Wolf, Rachel Blue, Mingyan Maggie Song, Jonathan D. Moreno, Guo-li Ming, and Hongjun Song. "Transplantation of Human Brain Organoids: Revisiting the Science and Ethics of Brain Chimeras." *Cell Stem Cell* 25, no. 4 (October 2019): 462-472. doi:10.1016/j.stem.2019.09.002.

Eiraku, Mototsugu, Kiichi Watanabe, Mami Matsuo-Takasaki, Masako Kawada, Shigenobu Yonemura, Michuru Matsumura, Takafumi Wataya, Ayaka Nishiyama, Keiko Muguruma, and Yoshiki Sasai. "Self-Organized Formation of Polarized Cortical Tissues from ESCs and Its Active Manipulation by Extrinsic Signals." *Cell Stem Cell* 3, no. 5 (November 2008): 519-532. doi:https://doi.org/10.1016/j.stem.2008.09.002.

Farahany, Nita A., Henry T. Greely, Steven Hyman, Christof Koch, Christine Grady, Sergiu P. Pasca, Nenad Sestan, et al. "The ethics of experimenting with human brain tissue." *Nature* 556 (2018): 429-432. doi:10.1038/d41586-

018-04813-x.

Gruen, Lori. "The Moral Status of Animals." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Last modified Fall 2017. https://plato.stanford.edu/archives/fall2017/entries/moral-animal.

Hostiuc, Sorin, Mugurel Constantin Rusu, Ionuţ Negoi, Paula Perlea, Bogdan Dorobanţu, and Eduard Drima. "The moral status of cerebral organoids." *Regenerative Therapy* 10 (June 2019): 118-122. doi:10.1016/j.reth.2019.02.003.

Hyun, Insoo., J.C. Scharf-Deering, and Jeantine E. Lunshof. "Ethical issues related to brain organoid research." *Brain Research* 1732 (April 2020): article #146653. doi:https://doi.org/10.1016/j.brainres.2020.146653.

Kadoshima, Taisuke, Hideya Sakaguchi, Tokushige Nakano, Mika Soen, Satoshi Ando, Mototsugu Eiraku, and Yoshiki Sasai. "Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell–derived neocortex." *Proceedings of the National Academy of Sciences* 110 (December 2013): 20284-20289. doi:10.1073/pnas.1315710110.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals.* Translated by Mary J. Gregor. Cambridge: Cambridge University Press, 1998.

Kant, Immanuel. *Groundwork for the Metaphysics of Morals.* Translated by Jonathan Bennett. (Unpublished manuscript, 2017), typescript. https://www.earlymoderntexts.com/assets/pdfs/kant1785.pdf.

Koplin, Julian J. and Julian Savulescu. "Moral Limits of Brain Organoid Research." *Journal of Law, Medicine, and Ethics* 47, no. 4 (December 2019): 760-767. doi:10.1177/1073110519897789.

Lagercrantz, Hugo and Jean-Pierre Changeux. "The Emergence of Human Consciousness: From Fetal to Neonatal Life." *Pediatric Research* 65, no. 3 (March 2009): 255-260. doi:10.1203/PDR.0b013e3181973b0d.

Lavazza, Andrea. "Human cerebral organoids and consciousness: a double-edged sword." *Monash Bioethics Review* 38, no. 2 (September 2020): 105-128. doi:10.1007/s40592-020-00116-y.

Lavazza, Andrea and Marcello Massimini. "Cerebral organoids: ethical issues and consciousness assessment." *Journal of Medical Ethics* 44 (September 2018): 606-610. doi:10.1136/medethics-2017-104555.

McMahan, Jeff. *The Ethics of Killing: Problems at the Margins of Life.* New York: Oxford University Press, 2002.

Muguruma, Keiko, Ayaka Nishiyama, Hideshi Kawakami, Kouichi Hashimoto, and Yoshiki Sakai. "Self-Organization of Polarized Cerebellar Tissue in 3D Culture of Human Pluripotent Stem Cells." *Cell Reports* 10, no. 4 (February 2015): 537-550. doi:https://doi.org/10.1016/j.celrep.2014.12.051.

Munsie, Megan, Insoo Hyun, and Jeremy Sugarman. "Ethical issues in human organoid and gastruloid research." *Development* 144 (March 2017): 942-945. doi:10.1242/dev.140111.

Nowakowski, Tomasz J., Alex A. Pollen, Elizabeth Di Lullo, Carmen Sandoval-Espinosa, Marina Bershteyn, and Arnold R. Kriegstein. "Expression Analysis Highlights AXL as a Candidate Zika Virus Entry Receptor in Neural Stem Cells." *Cell Stem Cell* 18, no. 5 (May 2016): 591-596. doi:10.1016/j.stem.2016.03.012.

Owen, Adrian. *Into the Gray Zone: A Neuroscientist Explores the Border Between Life and Death.* New York: Scribner, 2017.

Pera, Martin F. "Human embryo research and the 14-day rule." *Development* 144 (June 2017): 1923-1925. doi:10.1242/dev.151191.

Putnam, Hilary. *Reason, Truth, and History.* Cambridge: Cambridge University Press, 1981.

Rawls, John. *Political Liberalism.* New York: Columbia University Press, 1993.

Sakaguchi, Hideya, Taisuke Kadoshima, Mika Soen, Nobuhiro Narii, Yoshihito

Ishida, Masatoshi Ohgushi, Jun Takahashi, Mototsugu Eiraku, and Yoshiki Sakai. "Generation of functional hippocampal neurons from self-organizing human embryonic stem cell-derived dorsomedial telencephalic tissue." *Nature Communications* 6, no. 1 (November 2015): article #8896.

Sawai, Tsutomu, Hideya Sakaguchi, Elizabeth Thomas, Jun Takahashi, and Misao Fujita. "The Ethics of Cerebral Organoid Research: Being Conscious of Consciousness." *Stem Cell Reports* 13, no. 3 (September 2019): 440-447. doi:10.1016/j.stemcr.2019.08.003.

Shiraishi, Atsushi, Keiko Muguruma, and Yoshiki Sasai. "Generation of thalamic neurons from mouse embryonic stem cells." *Development* 144 (April 2017): 1211-1220. doi:10.1242/dev.144071.

Watanabe, Kiichi, Daisuke Kamiya, Ayaka Nishiyama, Tomoko Katayama, Satoshi Nozaki, Hiroshi Kawasaki, Yasuyoshi Watanabe, Kenji Mizuseki, and Yoshiki Sasai. "Directed differentiation of telencephalic precursors from embryonic stem cells." *Nature Neuroscience* 8, no. 3 (February 2005): 288-296. doi:10.1038/nn1402.

Computation and the Aprioricity of Mathematical Methods

Josh Katofsky

I. *Introduction.*

The prevailing view throughout the history of the philosophy of mathematics has been that mathematical discovery is an *a priori* process: we reach new mathematical knowledge independently from sense-experience, using only our logical faculties. While we may use our senses to problem-solve, a mathematical object in question has only been attained (or constructed, depending on your school of thought) if the mathematician reasoned to it independently from their sense-experience; this is the burden of a traditional mathematical proof. I will take this view as a baseline. But when computers enter the picture, does that change this epistemic relation? At what level of computer involvement in a proof is mathematical knowledge discovered by an *a posteriori* method: one in part constituted by reasoning external to the mathematician?

This paper investigates the effect of computation on the *methods* of mathematics, not on mathematical *objects*. This assumes that the aprioricity of a mathematical method is a distinct property from the *a priori* truth of a mathematical object, the latter pertaining to whether the mathematical knowledge itself requires sense-experience to be justified.[1] Rather, this paper will discuss the relationship between a *given method* used to attain a mathematical result and a *given mathematician*: did their internal reasoning constitute the entirety of the method, or has the computer, to some extent, done that work for them—work which they then had to *observe*? It is in this sense I could deem a mathematical method *a posteriori* regardless of whether the resulting knowledge can be in principle justified *a priori*; a method is *a priori* if and only if the origin of its deductions is in no way external to the mathematician, if any tools utilized by the mathematician could *in principle* be replaced by the mathematician.

In this paper, I first apply this criterion to argue against claims that the use of computation as a tool for brute-forcing in *computer-assisted proofs* makes said proofs *a posteriori* methods, making heavy use of the computer-assisted proof of the Four Colour Theorem as a case study. Then, I describe a famous unsolved conjecture which could indeed upset the *a priori* status of mathematical methods: the $P\overset{?}{=}NP$ problem. Namely, I argue that, by my criterion, the resolution of P=NP would make future mathematical methods *a posteriori*.[2]

II. *Computer-Assisted Proofs.*

(A) *Background.*

The term "computer-assisted proof" has been used to describe multiple types of computer involvement in mathematical methods, for example AI theorem-provers or the use of proof assistants. However, I will use the term in a narrow sense, only considering computer-assisted proofs-by-exhaustion, where a computer is used as a brute-forcing tool in a proof-by-cases with too many cases for a human to consider.

[1] This is contrary to the beliefs of thinkers like Imre Lakatos, who emphasized the sameness of discovery and justification.

[2] Though this question lies in the area of computational complexity theory, results in *computability theory*, such as undecidability, have also been applied to the epistemology of mathematics. However, the main results in this field have been limitative and thus do not have implications of the same type P=NP would.

Perhaps the most famous (and first) example of a computer-assisted proof is Wolfgang Haken and Kenneth Appel's 1976 proof of the Four Colour Theorem, which simply states that the maximum number of colours required to colour countries (or states, provinces, etc.) on any conceivable map so that no adjacent countries are the same colour will never exceed four.[3] Mathematically, it is a theorem in graph theory, stating that every planar graph (i.e., every graph that can be drawn in R^2 without edges crossing) has a chromatic number of four (i.e., the *minimum* number of colours required to colour all vertices, such that no adjacent vertices share a colour, is four).

Haken and Appel's technique, while made possible by modern computing advances, followed in the approach of Alfred Bray Kempe over a century prior in his failed attempt at the same result: they showed that it is impossible for all planar graphs *not* to be four-colourable. More specifically, they showed that any counter-example to the Four Colour Conjecture (i.e. a five-chromatic planar graph) must contain, as a subgraph, at least one of a so-called *unavoidable* set of graphs. They then used computation to prove that all such unavoidable graphs lead to contradictions by way of being *reducible*: they cannot appear in any five-chromatic planar graph.4 Accordingly, a five-chromatic planar graph could not exist, and the Four Colour Conjecture became a theorem.[5]

The proof took over 1,200 hours of computing time. The unavoidable set of reducible subgraphs was of size 1,936, with each subgraph requiring tens of thousands of computational steps to verify. This was the first prominent example of a computer-assisted proof. It is also an example of *unsurveyability*: the property of a

[3] Kenneth Appel and Wolfgang Haken, "The Solution of the Four-Color-Map Problem," *Scientific American* 237, no. 4 (October 1977): 108, www.jstor.org/stable/24953967.

[4] Ibid., 110-111. This was done with a process called *discharging*, an explanation of which is outside the scope of this paper.

[5] This provided a strict upper bound on the number of required colours for a planar graph due to the existence at the time of the (much more straightforward) *Five* Colour Theorem.

proof being so long that it is, for all intents and purposes, impossible for humans to read, rendering it impossible for humans to verify all of its steps. In fact, the referees of the *Illinois Journal of Mathematics* used a computer program itself to verify Haken and Appel's reducibility computations.[6] Initially met with some skepticism, this proof has gained increasing acceptance as it has been independently verified and simplified. To this day, no surveyable proof of the Four Colour Theorem has been found.

(B) *Computer-Assisted Proofs Are an* A Priori *Method.*

In any example of a computer exhausting cases of a proof, the Four Colour Theorem's proof being no exception, the reasoning contained within the proof was predetermined and known to the mathematician at the time the computation began; the mathematician possessed all requisite reasoning needed to write the program. Then, the computer, a fully deterministic machine when running brute-force algorithms, was used to carry out the mathematician's deductions.[7] The computer's assistance is merely an extension of reasoning internal to the mathematician, and as a result the computer-assisted proof is still an *a priori* method by my criterion.

(C) *Against Aposterioricity from Unsurveyability.*

It is worth noting that the above argument appears uncontroversial when the computer-assisted proof in question is *surveyable*. When a mathematician uses a computer to exhaust cases because they simply could not be bothered to do so manually, one cannot reasonably argue that the logic of the proof is in any way external

[6] Appel and Haken, "Four-Color-Map Problem," 121.

[7] In this paper, since we are only dealing with deterministic algorithms, I will assume that the steps taken by the computer correspond exactly to those in its code. In other words, I assume there are no *hardware errors* – for example stray photons causing bit flips in memory – as such errors are both unavoidable by the programmer and incredibly rare, making them philosophically equivalent to "hardware errors" in traditional mathematical methods, such as a writing utensil malfunctioning and deceiving a mathematician into proving something they did not mean to.

to the mathematician, for their reasoning was manifest in the computer program *and* we have a surveyable verification of this fact (although I will argue the latter is not a necessary condition for aprioricity). We would view surveyable computer-assistance as no different from an abacus or a compass in a traditional mathematical proof: a tool used by the mathematician to carry out their reasoning.

However, the most common argument against the aprioricity of computer-assisted proofs arises when, as with the Four Colour Theorem's, the proof is unsurveyable. If no mathematician can fully know what is happening within the proof and must rely on observing a program's output, is the reasoning that they possess not in some way incomplete? In this case, must the reasoning for the proof not fully originate from them? This sort of argument was most famously articulated by Thomas Tymoczko a few years after the proof of the Four Colour Theorem. As Tymoczko's general argument discusses epistemic qualities of both the proof and the result itself, it certainly envelops our question of the aprioricity of computer-assisted proofs as a method. He also makes statements that seem to directly engage with the status of methods, claiming that "this appeal to computer…is ultimately a report on a successful experiment. It helps establish the 4CT [Four Colour Theorem] (actually, the existence of a formal proof of the 4CT) on grounds that are in part empirical."[8] He concludes that while knowledge of unavoidability and reducibility helps us understand that, in principle, the exhaustion of cases *would* prove the Four Colour Theorem, the reason why this has successfully occurred is not known to the mathematician *a priori*: "it is not plausible to maintain that the 4CT is known by reason alone."[9] In sum, the claim is that, on account of unsurveyability, the relationship between the mathematician and their proof is no longer *a priori*.

The fact that the proof of the Four Colour Theorem has been surveyed by a *computer* will be unlikely to satisfy someone who holds this objection, as it merely

[8] Thomas Tymoczko, "The Four-Color Problem and Its Philosophical Significance," *Journal of Philosophy* 76, no. 2 (February 1979): 63, https://www.jstor.org/stable/2025976.

[9] Ibid., 77.

provokes the question of why we have *a priori* knowledge of steps carried out by the verifying computer program. It is also unsatisfactory to claim that the proof could be surveyed by a rational agent that can read inhumanly fast, or that perhaps lives for thousands of years longer than humans, since aposteriority from unsurveyability is predicated simply on proofs being known by the mathematician credited with them.[10] In a reality where mathematicians could survey formerly unsurveyable proofs, we would simply move the threshold for unsurveyability to exceed their new surveying capacity; this would not contradict the validity of aposteriority from unsurveyability.

However, I reject surveyability as a necessary condition for the aprioricity of a mathematical method. To find surveyability necessary presupposes that the computer-assisted part of the proof contains new reasoning of its own which needs surveying. Tymoczko is correct in stating that knowledge of unavoidability and reducibility does not constitute understanding of the proof; it is rather those facts *combined* with the surveyable and deterministic verification algorithm, written by Appel and Haken, that constitutes understanding. As McEvoy puts it, "…it is simply mistaken to say of a long proof that it is *a posteriori* simply on the basis of its length. What determines whether a proof is *a priori* is the type of inferential processes used to establish the conclusion of that proof."[11] When Appel and Haken observed the output of their computation, they did not need to survey the proof to understand the reason it had succeeded, as they had predetermined in their program exactly the conditions under which the computation would provide a given result. They possessed the knowledge to themselves manually complete each step of the proof; the only thing they lacked was adequate time.

As an analogy, take a computer program that determines whether or not

[10] Mark McEvoy, "The Epistemological Status of Computer-Assisted Proofs," *Philosophia Mathematica* 16, no. 3 (October 2008): 377-378, https://doi.org/10.1093/philmat/nkn014.
[11] Ibid., 380.

a number is prime by attempting all of its possible factors.[12] Even if one were to run that program on an input number so large that its number of possible factors is unsurveyable, such a program would still be an *a priori* method. Assuming one understands how factoring works (i.e., how the primality testing program works), they would understand in principle each step of the unsurveyable "proof" that a particular number is or is not prime. While, of course Haken and Appel's goal was much more complex than primality testing, I claim it is not philosophically different.

(D) *Against Aposterioricity from Errors.*

A related objection is that the computer introduces the possibility of errors to the proof. This unknown factor means that our understanding rests on "empirical assumptions about the nature of computers," namely, that all the code involved in the computation is perfectly written.[13] If that is not the case, it seems there is reasoning contained in the proof that does not reflect our *a priori* knowledge, rendering the method *a posteriori*.

If we grant that the reasoning of a computer-assisted proof is known *a priori* by the mathematicians and that the computer program is deterministic, any bugs in the code for a computer-assisted proof are evidence that our reasoning is sometimes misguided, not that computer-assisted proofs as a category are *a posteriori*. When conducting normal proofs, we err analogously. As is noted by E. R. Stewart and Israel Krakowski in both of their papers rebutting Tymoczko, we also create "bugs" when we express our mathematical reasoning using pencil and paper, abacuses, or any other of our methods for mathematics. In fact, Kempe's original failed proof of the Four Colour Theorem contained a "bug" that took 10 years to find.[14] [15]

[12] This example is not convoluted; for primality-testing we don't know methods much better than brute-forcing.

[13] Tymoczko, "Four-Color Problem," 77.

[14] E.R. Swart, "The Philosophical Significance of the Four-Color Problem," *American Mathematical Monthly* 87, no. 9 (1980): 703, https://doi.org/10.1080/00029890.1980.11995128.

[15] Israel Krakowski, "The Four Color Problem Reconsidered," *Philosophical Studies* 38, no. 1 (July

It should be noted that this argumentation extends to bugs in any code *used by* a mathematician, for example that of helper software or the computer's operating system. Errors in that software are analogous to one proving a mathematical theorem on the basis of another mathematical result that contains a logical mistake; we would not say that all proofs that rely on other results are then somehow *a posteriori*, simply that human error occurred somewhere along the chain of deductions for the proof in question.

Finally, if in response to this one claims that errors in computer-assisted proofs are philosophically different *only* because they are obscured by unsurveyability, their argument collapses to the argument from unsurveyability itself and thus cannot be sound if we grant the soundness of my argument in the previous section. We now turn to a demonstration of the level of computer involvement in a proof that would indeed upset its *a priori* status by my criterion.

III. *P=?NP.*

(A) *Background.*

Computational complexity theory is concerned with the classification of mathematical problems according to how difficult they are to solve by a computer. It formally investigates how, for a given mathematical problem, the number of steps its algorithm takes grows *in relation* to the size of the input.[16]

The complexity class P consists of problems that are fast to *solve*. Formally, this means that they take *polynomial time* to solve: the number of steps their solution algorithm takes, in the worst case, relative to an instance of the problem of an arbitrary size n, is some *polynomial* function of n.[17] For example, given two integers with n digits each, the multiplication algorithm we learn in grade school, whereby one it-

1980): 94, www.jstor.org/stable/4319399.

[16] This is specifically referring to time complexity, not space (i.e., memory) complexity, which escapes the scope of this paper.

[17] Stephen Cook, "The P Versus NP Problem," Clay Mathematics Institute, 1-2, www.claymath.org/sites/default/files/pvsnp.pdf.

erates over the digits of one number and multiplies them individually with the other number's, requires n^2 steps in the worst case.18 As n^2 is a polynomial expression, the problem of two-integer multiplication belongs to P.

The complexity class NP consists of problems for which solutions can at least be *verified* in polynomial time. By this definition, the P problems are a subset of the NP problems, as one can just re-solve a P problem to verify it in polynomial time. However, some problems in NP may take *exponential time* to solve.[19] Take the classic game of Sudoku and take n to represent the size of the sides of the board (and, by extension, the number of characters being used to fill the board). Here, the number of steps the algorithm must take grows in proportion to the *exponential* expression nn because it must permute all possible combinations of characters across all cells. This algorithm's number of steps grows *much* faster than a polynomial-time algorithm's does; solving a 20-by-20 Sudoku board requires up to 104 Septillion steps, which would not terminate by the time the sun envelops the earth (by contrast, multiplying two 20-digit numbers requires at most 400 steps and would finish in milliseconds on modern computers). However, crucially, given a finished Sudoku board, the algorithm to *check* if the answer is correct takes polynomial time; it consists of simply checking each square, which requires n^2 steps. This why NP includes Sudoku.[20]

This brings us to arguably the most consequential unsolved question in all of mathematics: the P=$^?$NP problem. It asks whether P is a *proper* subset of NP (that is, whether there are problems in NP that are not in P), or if, instead, all NP problems have an undiscovered polynomial-time solution algorithm. In other words, is it possible that every problem which is fast to verify is also fast to solve? We can naturally

[18] There are faster algorithms for multiplication that we can ignore without the loss of generality, as they are also polynomial time.

[19] Cook, "P Versus NP Problem," 2. Contrary to the reader's likely intuition, NP does not stand for "not polynomial". It stands for "non-deterministic polynomial," which has to do with the technical definition of an algorithm requiring greater than polynomial time.

[20] There exists a plurality of other complexity classes that escape the scope of this paper. For example, the problem of finding the best move in chess is in the class EXPTIME, where it takes exponential time to discover *and* verify a solution.

ask this question about *all* NP problems because of NP-*complete* problems which are at least as computationally difficult as every other NP problem; if a polynomial-time algorithm is found to solve *any* NP-complete problem, the algorithm will solve *all* NP problems in polynomial time and then P=NP.[21]

Most computational complexity theorists believe that P≠NP because of countless computer scientists' failure to find a polynomial-time algorithm for any of the *thousands* of NP-complete problems. There is also the possibility that a P=NP proof is non-constructive, establishing that polynomial time algorithms for NP problems exist but not providing us with them.[22] However, P=$^{?}$NP is still an open question with no present leads towards any resolution.[23]

From this point on, when I refer to P=NP, I assume a constructive result, meaning that NP problems become fast to solve in practice. For countless processes that currently require approximation techniques at best, or are fully intractable at worst, we could instead efficiently reach optimal solutions. Many such NP problems are of great practical interest. In fact, the problem of finding a proper colouring of a graph (planar or otherwise) is in NP and has myriad applications, for example in scheduling. Or, to draw on another aforementioned example, finding the factors of a number is an NP problem and is integral to the functioning of modern cryptography.

(B) ***P=NP and Mathematical Methods.***

If P=NP, perhaps no field would be changed more than mathematics itself. Most crucially, the process of finding a sub-length-n proof for a statement within a mathematical formal system (such as the current standard, Zermelo-Fraenkel

[21] Cook, "P Versus NP Problem," 4-5.

[22] Ibid., 7. While they may seem counter-intuitive, multiple non-constructive algorithm existence proofs have in fact been conducted.

[23] There is *also* the possibility that P=$^{?}$NP is independent of ZFC (i.e., neither provable nor disprovable within it), although it would be a statement of an entirely different variety from those that have thus far been shown to be independent.

set theory) is an NP problem; as n grows, the number of proofs to check grows exponentially, as they consist of exponentially many combinations of axioms and deductions. We currently know no polynomial-time algorithm to sift through these candidate proofs and instead need to brute-force through them, a process which is prohibitively slow for usefully large n (recall how Sudoku became intractable even for n=20). However, it is fast to *verify* that a given proof of length n is valid, simply by going over each of its n steps and checking that they are valid inference rules of the formal system.

As proof-finding is in NP, if P=NP, the search for proofs up to length n for a given mathematical statement would become fast.[24] The remarkable implication of this is best said by Kurt Gödel in his 1956 letter to John von Neumann:

> …the mental work of a mathematician concerning yes-or-no questions could be completely replaced by a machine. After all, one would simply have to choose the natural number n so large that when the machine does not deliver a result, it makes no sense to think more about the problem.[25]

One could provide the computer with a mathematical statement written in a formal language (say, a conjecture of great interest such as the Riemann Hypothesis), set n to an arbitrarily high number, and have the machine search for maximum length-n proofs of the statement within the formal system.[26] This could be done quickly even as n grows arbitrarily large, allowing us to set n so high that if a statement is prov-

[24] Walter Dean, "Computational Complexity Theory and the Philosophy of Mathematics," *Philosophia Mathematica* 27, no. 3 (October 2019): 417, https://doi.org/10.1093/philmat/nkz021.

[25] Kurt Gödel to John von Neumann, 20 March 1956. https://www.anilada.com/notes/godel-letter.pdf.

[26] We do not know how such an algorithm would search for proofs, only that it has to exist if P=NP. In fact, the intuition that there cannot exist a general, fast method to navigate exponentially large search spaces is a common argument for P≠NP.

able, we would with an incredibly high likelihood find a proof for it. It is here that "the P=$^?$NP question comes into contact with traditional foundational concerns in mathematics."[27] The intellect of mathematicians is only useful if it is impossibly slow for a computer to exhaustively search for proofs, where specialized mathematical knowledge and methods are needed to sift through the vast search space of mathematical statements.

(C) *P=NP Makes Mathematical Methods* A Posteriori.

From this point on, I will refer to the proof-finding algorithm that would exist in a P=NP world, described by Gödel in the previous section, as *Proof-Finder* and its proofs as *computer-generated proofs*. By my criterion, using Proof-Finder would be an *a posteriori* method of attaining mathematical knowledge, as the reasoning of a computer-generated proof would be *a posteriori* to the mathematician. The means by which a mathematician understands such a proof is *observing* the output of Proof-Finder, not generating it using reasoning internal to them. The deductive structure of a computer-assisted proof is understood by the mathematician before observing the output of the computation, while the deductive structure of a computer-generated proof is only known after observing it. In a P=NP world, any mathematician who could formalize the statement of the Four Colour Conjecture in a formal language could have discovered the proof of the statement without needing to understand unavoidability or reducibility.

There are some crucial points to specify about this claim. First, if my arguments in the previous section are sound, this claim holds whether or not a computer-generated proof is surveyable (Proof-Finder would certainly be able to generate proofs that are unsurveyable). This is because surveyability does not dictate the origin of the proof's reasoning; the origin is determined by analyzing the relationship between the mathematician's knowledge and the steps taken by the computer.

[27] Dean, "Computational Complexity Theory," 417.

Next, since Proof-Finder is using some formal system for mathematics, in isolation the proofs it generates are not any different from ones that a mathematician would construct in the same formal system. It bears repeating: the claim is not that computer-generated proofs rely on sense-experience to be justified, but that we could deem the proofs they generate valid without the mathematician themself having understood the structure of the proofs. This is because the reasoning does not come from the mathematician; it was located within the field of true statements of mathematics.

Finally, a corollary of the previous observation is that understanding of Proof-Finder's operations is fully distinct from understanding of its computer-generated proofs. Even if the mathematician has a perfect knowledge of Proof-Finder's inner workings, even if they are the one who proved that P=NP, they will only know the logic that allows Proof-Finder to be efficient but not necessarily that of the proofs it finds. Since computer-generated proofs would be part of an infinity of proofs within a formal system for mathematics, the very same formal system that grounds all traditionally-proven theorems, their deductions would have no relation to the logic of Proof-Finder. This stands in contrast to computer-assisted proofs, for which having written the proof-assisting algorithm necessarily entails an understanding of that proof's deductive steps because the computer only enters the picture once the mathematician knows the conjecture-specific logic of the proof.

IV. *Conclusion.*

In summary, my criterion for the aprioricity of a mathematical method—the origin of its deductive structure—was applied to two realms of mathematics, one historical and one hypothetical. The former, computer-assisted proofs by exhaustion, was shown to meet this criterion and I argued against claims that unsurveyability or the possibility of errors changed this fact. Then, I used the P$=^?$NP problem to illustrate the level of computer involvement in proofs that would make them *a posteriori* to the mathematician—namely, that the computer generates the reasoning behind the proof. I argued that this would occur in the scenario where P=NP.

My question was purposefully narrow in scope: how does the origin of a method's reasoning change when computation enters the picture? The intersection of computation and the philosophy of mathematics is a much wider area of inquiry. For future research, the impact of computer-assisted proofs, the P=$^{?}$NP problem, and other realms of computation on questions of mathematical justification, ontology, communication, and convincingness would all be fruitful topics.

Bibliography

Appel, Kenneth, and Wolfgang Haken. "The Solution of the Four-Color-Map Problem." *Scientific American* 237, no. 4 (October 1977): 108-121. www.jstor.org/stable/24953967.

Cook, Stephen. "The P Versus NP Problem." Clay Mathematics Institute, www.claymath.org/sites/default/files/pvsnp.pdf.

Dean, Walter. "Computational Complexity Theory and the Philosophy of Mathematics." *Philosophia Mathematica* 27, no. 3 (October 2019): 381-439. https://doi.org/10.1093/philmat/nkz021.

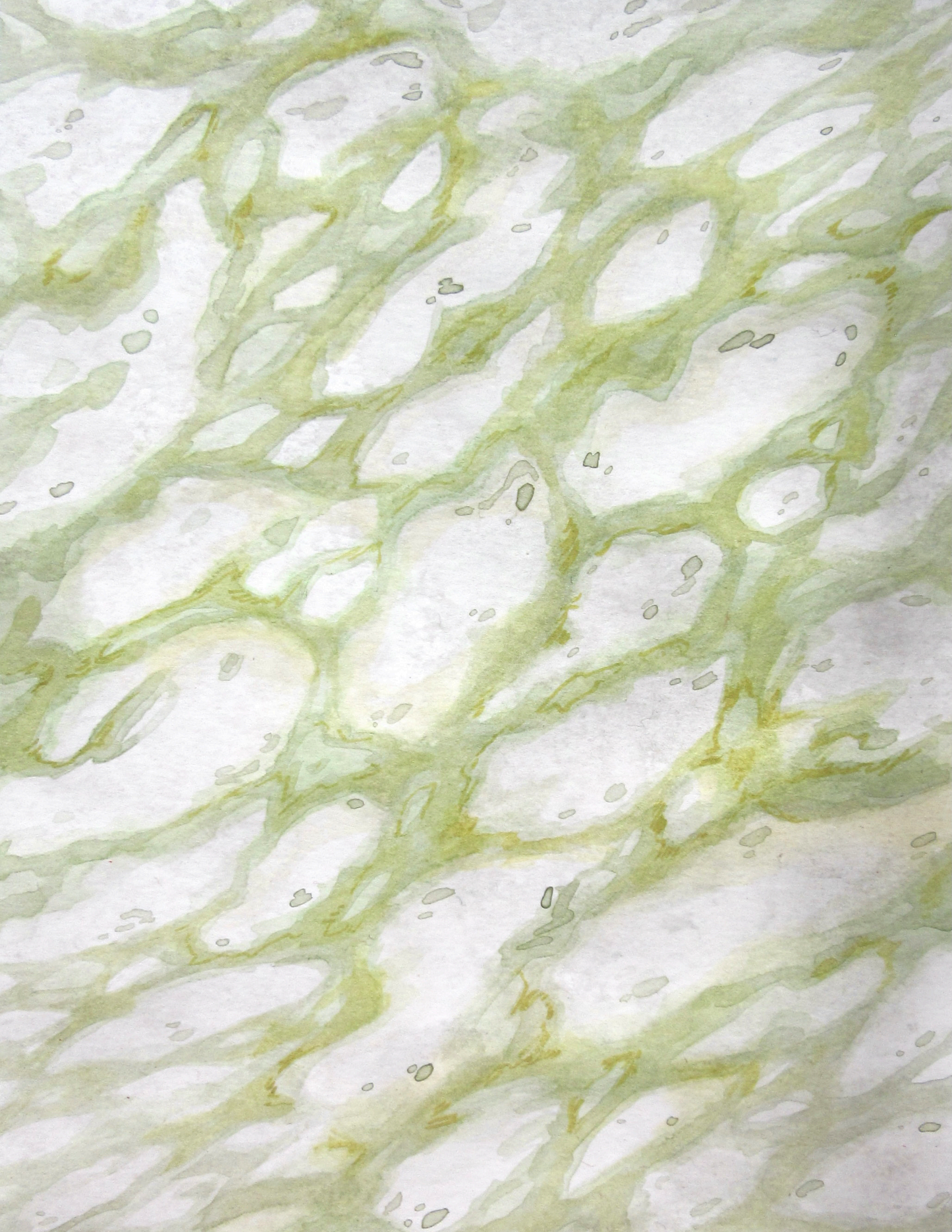
Gödel, Kurt. Kurt Gödel to John von Neumann, 20 March 1956. https://www.anilada.com/notes/godel-letter.pdf.

Krakowski, Israel. "The Four Color Problem Reconsidered." *Philosophical Studies* 38, no. 1 (July 1980): 91-96. www.jstor.org/stable/4319399.

McEvoy, Mark. "The Epistemological Status of Computer-Assisted Proofs." *Philosophia Mathematica* 16, no. 3 (October 2008): 374-387. https://doi.org/10.1093/philmat/nkn014.

Swart, E.R. "The Philosophical Implications of the Four-Color Problem." *American Mathematical Monthly* 87, no. 9 (1980): 697-707. https://doi.org/10.1080/00029890.1980.11995128.

Tymoczko, Thomas. "The Four-Color Problem and Its Philosophical Significance." *Journal of Philosophy* 76, no. 2 (February 1979): 57-83. https://www.jstor.org/stable/2025976.



Action as Addition: A Move Beyond Consequentializing

Pratik Mahajan

I. *Introduction.*

Act consequentialism holds the deontic properties (rightness or wrongness) of acts to be determined by the evaluative properties (goodness or badness) of their outcomes, i.e. what the actions bring about. The deontological position, by contrast, is that there are moral constraints against performing certain acts, irrespective of how much goodness or badness those actions may produce. The act of killing is one such action, and the deontological position is that it is always impermissible to kill, even to prevent more killings.

In Part II of this essay, I will outline the consequentializing project, which argues that constraints against killing can be accommodated by the consequentialist framework. In Part III, by engaging with Daniel Muñoz's argument, I will explain why the consequentializing project is doomed to fail from the start. In Part IV, I will consider the novel concept of Action as Addition, as well as its implications for the consequentializing moral theorist. In Part V, I will argue that Action as Addition gives us a non-deontological and a non-consequentialist principle—*an agent must act to maximize the good or minimize the bad added by that action to an outcome*—that can accommodate an absolute constraint against killing. Finally, in Part VI, I will argue against the objection that such a principle would permit agents to be bystanders who would fail to prevent other agents from being harmed or causing harm to others.

II. *The Consequentializing Project.*

Consequentialism as a moral theory asks agents to maximize the goodness of the outcomes they produce. At times the results of this view seem to conflict with common moral intuitions. As a result, the consequentializers take on consequentialism and attempt to show how it can accommodate these common moral intuitions, one being an absolute constraint against killing.

I will elucidate the consequentialiser's argument through Judith Thomson's *Footbridge* scenario. A trolley is moving on a track in the direction of five innocent workers. You happen to be on a footbridge above the track and can stop the trolley by pushing a large man off the footbridge down onto the track between the trolley and the five workers, killing the large man to save the five. You therefore have two options: kill the large man to save the five, or don't kill the large man and let the five die.[1] The consequentialiser's strategy to come to the judgement that you *should not* kill the large man is to claim that the outcome that contains an intentional killing is intrinsically bad. To intentionally kill the large man would be worse than to let the five die in an accident.

However, the consequentialiser is forced to update their strategy in a scenario such as *Villainous Footbridge*, in which a villain has set the trolley in the direction of the five workers with the intention of killing them, resulting in a situation where five killings are set against one killing.[2] Here, the most prominent strategy is to include 'evaluator relativism' within act consequentialism, in which the goodness of an outcome can vary depending on the position occupied by the evaluator.[3] Thus, the re-

[1] Judith Jarvis Thomson, "The trolley problem," *Yale Law Journal* 94.6 (1985): 1409.

[2] Daniel Muñoz, "The Rejection of Consequentialising," (unpublished manuscript), typescript, 5.

[3] Douglas W. Portmore, "Combining teleological ethics with evaluator relativism: A promising result," *Pacific Philosophical Quarterly* 86.1 (2005): 96.

sulting Non-Egoistic Agent-Relative Consequentialist (NARC) strategy in *Villainous Footbridge* is to allow *you* as the evaluator to argue that an outcome in which *you* kill is worse than an outcome in which the villain kills. However, in another scenario called *Redemption Footbridge*, where *you* are the villain who is now faced with the choice of killing the large man to prevent yourself from killing five, evaluator relativism is no longer relevant.[4] The consequentializer may then argue for the inclusion of 'temporal indexing' within consequentialism to come to the verdict that the agent must consider the killings at a particular moment in time.[5] You as the villain are not permitted to set into motion the act of killing at the particular time that you are considering whether to push the large man or not. Thus, the most prominent consequentializing strategy includes holding the outcome in which an intentional killing occurs as intrinsically bad, indexed to the evaluator and to time. How plausible is this strategy?

III. ***Consequentializing Is Destined to Fail.***

At the heart of the consequentializer's strategy is to make the distinction between *killing someone* and *letting someone die.* In *Footbridge*, by not killing the large man, you don't also *kill* the five on the tracks; you simply *let them die*. In criticizing this step, Daniel Muñoz has argued that consequentializers cannot make this distinction without giving up act consequentialism's 'Compelling Idea.' The Compelling Idea—*that one is always permitted to act to maximize the goodness of outcomes*—is only true, Muñoz argues, if performing an action is understood as the production of outcomes. This concept, termed 'Action as Production,' is essential to consequentialism's Compelling Idea because to understand actions as anything but outcome-producing is to give up the consequentialist claim that the deontic properties of actions are determined by the evaluative properties of outcomes, instead of other reasons such as conforming to moral norms.[6] Given that Action as Production is central to consequentialism,

[4] Muñoz, "Rejection of Consequentialising," 5.
[5] Richard Brook, "Agency and Morality," *The Journal of Philosophy* 88.4 (1991): 198.
[6] Muñoz, "Rejection of Consequentialising," 11.

Muñoz argues that, when the consequentializer makes the distinction between the actions of killing someone and letting them die, they are no longer operating within the framework of Action as Production.

The consequentializer cannot claim that your decision not to push the large man off the bridge is a decision to *let the five die* and not that of *killing the five.* Irrespective of whether the term 'letting die' or 'killing' is used, the same outcome is realised—*the five workers die*. If Action as Production is right, argues Muñoz, then letting the five die and killing the five are not two distinct acts, but the *same* act by virtue of producing the same outcomes.[7] The formidable implication of Muñoz's argument is that it denies consequentializers their strategy of distinguishing between killing and letting die. Without this strategy, a consequentializer cannot accommodate a constraint against killing even in the original *Footbridge* scenario. Therefore, the consequentializing project is destined to fail from the beginning, even before evaluator relativism and temporal indexing enter their strategy.

How might a consequentializer respond to Muñoz denying them the strategy of distinguishing between killing and letting die? The consequentializer may modify the concept of action, arguing for actions to be something other than outcome-producing. However, given that Action as Production is a necessary component of consequentialism's Compelling Idea, the consequentializer will need to further modify the Compelling Idea itself. Is such a two-fold project possible? I argue that it is possible, but that it cannot stay true to consequentialism.

IV. ***From Action as Production to Action as Addition.***

To begin the twofold project, my first claim is that when we perform an action, we don't bring about an outcome—we simply add a specific *constituent part* to the outcome that is eventually brought about by the performance of multiple actions. In other words, we have the new concept of Action as Addition—*what it means to perform*

[7] Ibid., 12.

a particular action is to add a specific part to the overall outcome. But what exactly is a *part*, and what does it mean to *add* one to the outcome?

Consider the *Footbridge* scenario again, and the outcome in which you choose to push the large man. This outcome can be divided into parts determinable as having been added through specific actions performed by multiple agents. For instance, the trolley followed a particular path of the track because of the passengers' desired destination, the five workers were tasked to work on the tracks possibly under orders of a maintenance authority, and the large man was killed by you through being pushed from the footbridge. Only when all these actions are performed, or when all their parts are added, does the outcome come into existence. Thus, a *part* of an outcome is the effect of a specific action, and it is *added* to the outcome through being performed.

My second claim is that the agent is responsible only for the part added to the outcome by their action. When you choose to push the large man off the footbridge and thereby kill him, you are responsible for adding that constituent part to the outcome. The argument for this claim bears a similarity to the consequentialist's argument, as according to both an agent is responsible for the consequences of their actions by virtue of having decided to perform them. However, under the principle of Action as Addition, actions add constituent parts to outcomes. Thus, agents are responsible only for the addition of parts resulting from their own actions, and not for the outcome as a whole.

Any moral theorist who adopts these two claims has already moved away considerably from consequentialism. It is important to remind ourselves that, even prior to the Compelling Idea, the consequentialist claims that deontic properties of actions are determined by the evaluative properties of outcomes. However, if Action as Addition is true, then any moral theory which adopts it must hold that the deontic properties of actions are determined by the evaluative properties of *the parts* added to the outcomes, instead of the outcomes themselves. We can expect the consequentializer to protest at this move away from consequentialism, which has begun even before we modify the Compelling Idea. Should this move disappoint the consequentializing moral theorist?

I argue that it should not. Moral theorists attempting to consequentialize constraints want to argue in favour of constraints against killing while denying that facts about the action itself can determine the deontic properties of actions. These moral theorists only need to show that the deontic properties of actions are determined *not* by facts about *what the action itself is*, but by facts about *what the action does*. The only notion of *what an action does* contained in the principles of consequentialism is that it brings about outcomes. However, by regarding action as addition and not production, a moral theorist who wants to reject the deontological argument while keeping constraints—such as those against killing—can do so without holding on to consequentialism. If what actions do is *add constituent parts to outcomes*, then the previously consequentializing moral theorist can move away from consequentialism and accept a different ethical theory that holds the claim mentioned above: *the deontic properties of actions are determined by the evaluative properties of the parts added to the outcome*.

But what would make such an ethical theory based on Action as Addition so compelling that the consequentializer would give up consequentialism's original Compelling Idea? If the rightness or wrongness of an action is to be determined by the goodness of that action's specific part, then we may construct the following principle: *an agent must act to maximize the good or minimize the bad added by that action to an outcome.* Thus, we have a central principle that is consistent with common-sense rationality and directs agents to maximize the good that their actions can add.

Note here that this principle cannot be adopted by the consequentialist, because the consequentialist is committed to evaluating the entire set of outcomes produced by actions. By contrast, my claim is that the rightness or wrongness of actions is determined not by their outcomes at all, but by the *part* added by that particular action to an outcome. How successful is this non-deontological—*and yet non-consequentialist*—idea in accommodating a constraint against killing?

V. *Killing: An Intrinsically Bad Part to Add.*

If the new principle is to accommodate an absolute constraint against killing, it must mean that the part that contains a killing has significant disvalue. Now, reconsider *Footbridge.* I have argued that if you choose to push the large man off the footbridge, the part that *you* add to the outcome is that of killing the large man, which has significant disvalue. You are *exclusively responsible* for that part's addition, and not for the other parts. However, to refrain from pushing the large man is merely to refrain from adding a part to the outcome, thus constituting a permissible action. The fact that the trolley will accidentally run over the five workers is unrelated to your action. Thus, the action that you are required to perform in *Footbridge* is to refrain from adding the killing of the large man to the outcome. This is consistent with the principle that you as an agent ought to add the maximum good to the outcome.

A similar verdict is the outcome of applying Action as Addition to *Villainous Footbridge.* While the five workers are intentionally killed by the villain, that part of the outcome was not added through your action. *You* as an agent add a bad part to an outcome if you kill the large man, but you when you refrain from killing you add nothing. By refraining, you continue to satisfy your duty to minimize the badness *your* actions can add to an outcome. Thus, *you* are required to refrain from killing.

For some readers, this line of reasoning will at first appear controversial. If an agent is only responsible for what they add to outcomes, then a grave implication would be that agents no longer need to consider themselves responsible for intervening to stop bad additions by other agents. This is an important objection with implications beyond merely the issue of constraints against killing. I will reply to this wider objection in Part VI. For now, let us consider the final scenario.

In *Redemption Footbridge,* you are the villain who has intentionally performed the action of adding the part in which the trolley is headed towards the five workers to kill them. *Surely now, acting within the framework of Action as Addition, we must conclude that you are permitted to kill the large man to minimize the badness you add to the outcome!* However, I argue that this is not the case—even here an agent would be required by the new principle to refrain from killing the large man. The principle permits the agent to perform actions if, out of the range of alternatives available to them, the chosen course of action *adds* the maximum of goodness or the minimum of badness to the outcome. In *Redemption Footbridge*, two distinct actions add two separate parts to the outcome, albeit by the same agent. You sending the trolley in the direction of five workers with the intention of killing them is an impermissible part that you are responsible for adding. However, when considering the next course of action, whether to kill the large man or not, you continue to be required to refrain from adding the part where the large man is killed, because this part also contains a killing. *Every action adds a specific part to an outcome*, and thus the deontic status of that act must be judged only on the basis of the goodness of that part which the action adds. The part resulting from your killing of five workers has already been added, and so an impermissible action has thereby already been performed. Performing the action of adding the large man's killing to the outcome continues to be impermissible because it adds a bad part to the outcome.

I expect this line of reasoning to also not be acceptable to some readers. They may protest that you as an agent have not yet added the part where the five workers are killed; *the killings can still be prevented*, albeit by performing another killing. In fact, a similar objection was raised by Howard in his response to Setiya's project of agent-neutral consequentializing of constraints against killing. According to Howard, Setiya's argument that killing one person to prevent five further killings is worse than five random killings overlooks the fact that "all the ethical damage has not been done prior to the murder of the five."[8] Would a similar objection become problematic for my non-deontological and non-consequentialist principle's verdict in *Redemption Footbridge*? I argue that it would not.

I concede that if the part of the outcome in which the five workers are run over by the trolley that *you* send in their direction has not yet been actualised, then the ethical damage can still be undone. But what does it mean to *undo* a bad part that an action is about to add to the outcome? Undoing the ethical damage added by an action is equivalent to adding a *new constituent part* that cancels out the addition of a previously added part. This new, separate action that you could perform (pushing the large man) continues to be restricted by the maxim of minimizing the badness of the part that this specific action adds to the outcome. Hence, you continue to be required to refrain from performing the act of killing the large man.

This is not to say that as an agent you are not required to undo the ethical damage that your action is about to add. The claim only is that the action you could perform to undo the ethical damage is itself an impermissible one. Thus, if none out of the courses of action that you are permitted to perform can undo the ethical damage that your previous action is going to add, then as an agent you must accept that your previous action *cannot be undone* and take the moral blame for having performed an impermissible act in the first place.

VI. *Against Being a Bystander.*

It could be thought that a principle which guides agents to maximize only the good that *their* actions add to outcomes could encourage them to be passive bystanders. Common moral intuitions ttell us that one should—when possible—intervene and stop other agents from being harmed, whether it is due to an accident or be-

[8] Christopher Howard, "Consequentialism and Constraints," (unpublished manuscript, 2020), typescript, 10.

cause of other agents intending harm through their actions. To hold only the goodness of *one's own* additions to be of moral significance is selfish at best and morally reprehensible at worst. It may appear to some readers that this is exactly what my principle allows, but I want to reassure them that the principle would almost always prevent agents from acting selfishly and becoming bystanders.

Note that my claim so far has been that we are responsible for the actions we choose to perform, and therefore the parts that we add to outcomes. However, the principle may require certain additions by agents to maximize the goodness they can add to outcomes irrespective of whether the agent has a desire to do so. Consider Singer's *Drowning Child* scenario. You see a child drowning in a pond and rescue her without any physical harm to yourself. However, by getting in the pond you will get your clothes dirty.[9]

Now, *you* were not the one who added the part in which the child drowns in the pond, and so you are not responsible for it. However, considering that rescuing the child does not involve the addition of a constitutive part that has significant disvalue, out of the available courses of action the one in which you rescue the child adds the most goodness to the outcome. Getting your clothes dirty does not contain significant disvalue, and so there is no moral basis upon which to choose not to rescue the child. On the other hand, in *Footbridge* the addition of the large man's killing contains significant disvalue, restricting you from preventing the five workers' deaths. Thus, if you continue to be a bystander in *Drowning Child*, you fail to perform that action by which the maximum good can be added to the outcome.

VII. Conclusion.

This paper's aim has been to show how consequentializers can give up consequentialism to accommodate constraints against killing but still reject the deontolog-

[9] Peter Singer, "The drowning child and the expanding circle," *New Internationalist,* 5 April 1997, 122. https://newint.org/features/1997/04/05/peter-singer-drowning-child-new-internationalist.

ical position. Such a move employs the concept of Action as Addition, which posits the compelling principle that it is always permissible for an agent to act to maximize the goodness and minimize the badness added by that action to the outcome. This concept of action, according to which actions (ethically speaking) do nothing but add parts to outcomes, can accommodate an absolute constraint against killing. While an agent remains responsible only for the parts that their actions add to outcomes, the agent is also responsible if they fail to maximize the goodness of the part they add, thereby preventing them from being a bystander in cases that do not involve them killing. Even if the principle of Action as Addition is coherent, an additional theory of value is needed to provide moral guidance to agents. Nevertheless, the upshot of this essay is that *at least* absolute constraints against killing can be accommodated from a non-deontological position.

Bibliography

Brook, Richard. "Agency and Morality." *The Journal of Philosophy* 88.4 (1991): 190-212.

Howard, Christopher. "Consequentialism and Constraints." (unpublished manuscript, 2020), typescript.

Muñoz, Daniel. "The Rejection of Consequentialising." (unpublished manuscript), typescript.

Portmore, Douglas W. "Combining teleological ethics with evaluator relativism: A promising result." *Pacific Philosophical Quarterly* 86.1 (2005): 95-113.

Singer, Peter. "The drowning child and the expanding circle." *New Internationalist,* 5 April 1997. https://newint.org/features/1997/04/05/peter-singer-drowning-child-new-internationalist.

Smart, J.J.C. and Bernard Williams. *Utilitarianism: For and against.* Cambridge University Press, 1973.

Thomson, Judith Jarvis. "The trolley problem." *Yale Law Journal* 94.6 (1985): 1395-1415.