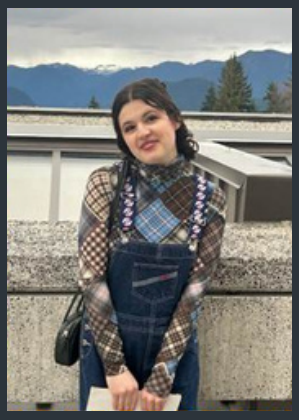
SIMON FRASER UNIVERSITY

2026
ISSUE 2

PHILOSOPHY NOUVEAU

TABLE OF CONTENTS

meet the team:



LUCIA PISTRIN:
co-chief journal editor

BLAKE ROBERTSON:
co-chief journal editor

GWENDOLYN GONZALEZ:
associate editor

RAMANJIT SAHOTA:
marketing coordinator

An Analysis of Fischer and Ravizza's View of Moderate Reasons-Responsiveness

Written by: Diana Chircea

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau
Simon Fraser University
Accepted February 28, 2026. Published March 30, 2026

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau

Abstract

Control is what grounds moral responsibility, often understood as being the ability to do otherwise. However, this view of control fails within a deterministic world wherein agents do not have the ability to choose their actions. In Responsibility and Control, Fischer and Ravizza argue against this traditional view of control, in favor of one compatible with determinism. This compatibilist form of control, called guidance control, is dependent on two factors: ownership of the mechanism which issues the action, and the reasons-responsiveness of said mechanism. While both aspects are conjunctively needed to ground moral responsibility, I argue that reasons-responsiveness should not be considered a sufficient condition for moral responsibility.

Simon Fraser University
Accepted: February 28, 2026. Published: March 30, 2026.

Introduction

Control is what grounds moral responsibility, often understood as being the ability to do otherwise. However, this view of control fails within a deterministic world wherein agents do not have the ability to choose their actions – they are unable to do otherwise. In Fischer and Ravizza's book, Responsibility and Control: A Theory of Moral Responsibility, Fischer and Ravizza argue against this traditional view of control, in favour of one that is compatible with determinism. This compatibilist form of control, called guidance control, is dependent on two factors that work in conjunction with one another: (1) ownership of the mechanism which issues the action, and (2) the reasons-responsiveness of said mechanism. While both aspects are needed to ground moral responsibility, I argue that reasons-responsiveness should not be considered a sufficient condition. To demonstrate my point, I will begin by providing a thorough explanation of guidance control and its parts. Next, I will outline an objection using a counterexample, wherein an agent meets the criteria for guidance control and yet intuitively does not appear to be morally responsible. Afterwards, I will provide an objection to my proposed view with the aid of Coates and Swenson's framework, in which they introduce a gradable aspect to Fischer and Ravizza's original view of reasons-responsiveness. I will conclude by providing a potential response to the objection, hopefully having shown that reasons-responsiveness may not be sufficient for grounding moral responsibility.

Guidance Control and Reasons-Responsiveness

The first requirement of guidance control for moral responsibility is that an agent takes ownership of the mechanism that produces the action. This mechanism could include practical reasoning, impulse, and habit, among others. Fischer and Ravizza argue that "what is relevant to the agent's moral responsibility is the actual-sequence mechanism." (444) This actual-sequence mechanism is the mechanism that actually produces the action and thus the mechanism that the

agent takes ownership of. For an agent to take ownership of a mechanism, they must take responsibility for it as well as the actions it issues.

According to Fischer and Ravizza, there are two contexts – the non-reflective case and the reflective case – in which an agent can take ownership of their mechanism; both are considered historical processes. The non-reflective case involves agents acquiring dispositional beliefs about themselves as one whose actions have consequences and who are subject to reactive attitudes like as punishment or praise. Note that these dispositional beliefs must be acquired in an appropriate way from education, or life experience, and not through coercion. In the non-reflective case, an agent can be said to have taken ownership of their mechanism without needing to engage in philosophical reflection. The reflective case is somewhat similar, with the difference being that it addresses agents who do engage in philosophical reflection and question whether they truly can be praiseworthy or blameworthy. However, Fischer and Ravizza point out that "practical purposes" may persuade these reflective agents to once again accept that they are apt candidates for reactive attitudes, taking ownership of their mechanisms once more. (442)

The ownership criterion for guidance control is a means of ensuring that moral responsibility is not attributed to agents who are manipulated or forced into doing an action. Consider the example of John, who, on account of a device implanted in his head, gatecrashes whenever he gets the opportunity. John, being unaware of this device, cannot take ownership of it or any of the actions it issues, barring him from moral judgment. Even if John was aware of the device, this alone would still not suffice unless he had taken responsibility for it through historically endorsing it as a part of his agency.

The second requirement for guidance control is the reasons-responsiveness of the mechanism. This criterion is how Fischer and Ravizza bypass control's requirement for an agent to have the ability to do otherwise. Instead of grounding

control in the ability to do otherwise in the current world, they test if an agent's mechanism has the ability to recognize and react to sufficient reasons in possible scenarios. In other words, they test how responsive the mechanism is to reasons in possible worlds. This responsiveness itself consists of two aspects: reactivity and receptivity. The relationship between the two is asymmetrical in the view of moderate reasons-responsiveness (MRR) for which Fischer and Ravizza argue. That is to say that under MRR, reactivity to reasons needs to be weak, but receptivity needs to be regular. Reactivity can be understood as an agent's ability to transfer reasons into choices. Receptivity can be understood as an agent's ability to recognize that sufficient reasons exist.

As a means of demonstrating how reactivity and receptivity operate, consider the example of James, who is debating if he should buy tickets to the upcoming Sabrina Carpenter concert. James acknowledges that he would have to miss an extremely important meeting to see Sabrina, but because he is such a big fan of hers, he is willing to miss the meeting and go. In fact, the only scenario in which James would miss the Sabrina concert would be if his boss threatened to fire him if he missed work. This example is meant to show that James' actual-sequence mechanism is at least weakly reactive on account of there being at least one scenario in which he would miss the Sabrina concert. The reason reactivity can remain weak in comparison to receptivity is because it "is all of a piece." (Fischer and Ravizza 73) Essentially, as long as a mechanism can react to a sufficient reason to do otherwise, then it can react to any sufficient reason to do otherwise; meaning that it could have reacted differently to the actual reason to do otherwise. So, James not going to the concert if his job was at risk, shows that his mechanism is generally able to react to reasons to do otherwise. Thus, he would have been able to react to the actual reason not to go - that he would miss the important meeting. Moreover, it is because he did not react appropriately to that reason, that he can be held morally responsible.

As opposed to reactivity, receptivity needs to be regular. Consider James once more, except this time he is considering the price of the tickets. He recognizes that if the tickets were $300, then there is a sufficient reason to not buy them. As long as James' mechanism now reacts to the recognition of this reason – given that his mechanism is also weakly reactive – then he is morally responsible. However, imagine that James only understood the $300 price as a sufficient reason to not go. If the tickets were $400, $500, or even $1000, he would not view those new prices as sufficient reasons to not go to the concert, demonstrating that his mechanism is only weakly receptive. Certainly in this scenario we would question whether he is genuinely morally responsible. From here, it's clear that the criteria for receptivity cannot merely be that an agent's mechanism can recognize a sufficient reason just in one instance; instead, a mechanism must show a regular and understandable pattern of recognition. We need to be able to see that the mechanism shows a regular and understandable pattern of recognition. In this case, 'understandable' can be simply understood as the reasons relating to each other coherently as well as relating to the objective. If James' goal is to save money, for his mechanism to have an understandable pattern of reason recognition, he would need to recognize both the $300 price and the $400 price as sufficient reasons to not go to the concert. The pattern itself can simply be understood as what reasons the agent would find sufficient given different circumstances. For example, it would test other scenarios where James might find the price of $300 a sufficient reason to not buy something. For example, James might find a price of 300$ a sufficient reason to not buy a T-shirt, but maybe not for a Metallica concert.

Additionally, a pattern must be at least minimally grounded in reality. James cannot say he finds the price of $300 a sufficient reason to not buy a ticket if James has no money. Further, it should be noted that for an agent to be found morally responsible, they must be receptive to moral reasons. According to Fischer and Ravizza, moral reasons are simply reasons that stem from the balancing of an individual's own interests against the interests of those around

them. (76) For James, this would mean that he should be able to recognize that he has a duty to his colleagues to do his job and go to the important meeting. That is all to say that a receptive mechanism needs to be regular, follow an understandable pattern, which is minimally grounded in reality, and recognize at least some moral reasons.

Counterexample

As an example, let us consider Leo. Leo is a businessman, who has business meetings. Leo typically accompanies his mom to her appointments. He knew of his mother's upcoming doctor's appointment in advance, and yet he scheduled an important meeting during that time. His mother only wants him there for support; she is self-sufficient and has gone by herself before with no issue. In fact, it is only recently that she has started asking Leo to accompany her. Leo knows that this is the only time he has ever missed an appointment which she asked him to attend, and also knows that it causes her great distress. However, this was the only time the client could have the meeting. Leo knows that this doctor's appointment is not important, it is a simple prescription refill, and does not consider his mother's request as a sufficient reason to miss this meeting. Let us stipulate that Leo takes ownership of the mechanism that produced this action. Additionally, let us say that it is both weakly reactive and regularly receptive. For example, if this was a very important appointment for his mother, he would have cancelled with the client and gone with her. Further, he does exhibit an understandable pattern of reasons that is grounded in reality and takes moral reasons into consideration. For instance, he once missed his son's soccer game because he could not reschedule a meeting with a client. In fact, whenever given the choice between a meeting that cannot be rescheduled and a family matter, he will always pick the meeting.

Yet, despite his being both regularly receptive and weakly reactive to reasons, it does not appear that he is intuitively morally responsible. This intuition may stem

from three main considerations: first, the moral stakes are overstated; second, his actions are not demonstrative of a pattern of neglect; third, there appears to be an aspect of moral luck involved. Beginning with the first issue, it appears that his mother may be overstating the moral stakes. Her distress, while it may be genuine, is arguably disproportionate given both her independence and the unimportance of this doctor's appointment. Consider further that if she were not distressed, there would not be any moral failure on Leo's part. This shift in whether Leo is subject to moral judgment indicates that his mother's distress is not a robust moral reason; instead, it is a contingent emotional reaction.

The second possible reason for this intuition could be that Leo missing the appointment is not a common occurrence. Recall that while moral responsibility does require the mechanism to have an understandable pattern of reason recognition, it does not require a pattern of reactive failures. In other words, it does not require Leo to repeatedly neglect his mother for him to be morally responsible. While he does showcase a pattern – when there is an unavoidable scheduling conflict, he prioritizes work – he still consistently fulfills his familial duties.

Lastly, the moral luck involved in whether he is morally judged may aid the intuition that he is not morally responsible. Once again, consider that he is only morally judged if his mother reacts negatively – a factor completely outside the scope of his control. It is true that Fischer and Ravizza place an emphasis on an agent's mechanism and not the outcomes of his actions, yet this scenario seems to hint at a tension: if moral responsibility hinges entirely on an agent's internal mechanisms, it overlooks the external factors like the responses of the individuals around them. While Leo's mechanism is responsive, it appears to be insufficient to warrant blame, suggesting that reasons-responsiveness is not sufficient to ground moral responsibility.

Coates and Swenson Objection

The extension of Fischer and Ravizza's original framework by Coates and Swenson in "Reasons-Responsiveness and Degrees of Responsibility" provides a substantial objection to my counterexample. Their extension introduces gradability to moral responsibility judged by the proximity of differing possible worlds. Here it should be noted that their view, despite adding gradability, is still a threshold view that requires "agents [to] satisfy some minimal threshold for moral responsibility." (Coates and Swenson 630) For now, consider the following example as a means of understanding how the authors extend the MRR.

Tina promised to pick Ted up from the airport. However, Tina is clinically depressed, and as such, fails to pick him up, opting to stay in her apartment. She later tells Ted that she was aware that she had an obligation to pick him up, but was simply unable to get out of her bed. The next time Ted needed to be picked up, he asked Ed for help. However, Ed also failed to help Ted because there was an interesting new show playing on Netflix, also opting to stay in his apartment. Ed also tells Ted that he was aware that he had an obligation to pick him up, but simply did not find it a priority.

In this scenario, both Ed and Tina meet the requirements for guidance control. They are both weakly reactive; if there was a fire, they would have left their apartment, or if they were offered $1000, they would have immediately left to pick up Ted. Additionally, their recognition of the obligation they had to pick up Ted implies that they are both regularly reasons-receptive. So then the challenge becomes attempting to vindicate the distinction between the two, why Tina is intuitively less morally responsible than Ed.

Coates and Swenson propose looking at the scenario and comparing the nearest possible worlds in which Tina and Ed react differently. They claim that the nearer the possible world in which an agent reacts differently, the greater the degree of

their responsibility. (636) In this case, “nearness” to the actual world is measured by how many factors would be different. For reference, consider world A and world B, which are nearly identical with one exception: world A has a tree with 236 leaves, and that same tree in world B has 235. These two worlds would be considered extremely close as a result of the only difference between them being nearly negligible. Now, once again looking at Ed and Tina, it would appear that the world in which Ed goes to pick up Ted is much closer than the world in which Tina goes to pick up Ted. This is because all that would need to be changed for Ed to pick up Ted would be him not seeing that a new show was playing. In contrast, for Tina to pick up Ted, many more factors would need to be changed, including factors which influenced Tina's depression and how it manifests itself. Therefore, since the possible world in which Ed picks up Ted is closer than that in which Tina picks up Ted, he is more morally responsible than she is.

Applying Coates and Swenson’s extension of MRR to Leo, we may be able to say that while he may not have been morally responsible under the binary view originally postulated by Fischer and Ravizza, he certainly can be considered at least partially responsible under the new view that introduces gradability to moral responsibility. This consideration directly challenges the original intuition that was proposed. Consider the three arguments that I made in support of the intuition that Leo is not morally responsible: that the moral stakes are overstated, that this was an isolated incident, and that the scenario is susceptible to moral luck. Starting with the first argument, while it may be true that his mother’s distress is disproportionate, under this new view, her reaction no longer needs to completely wipe away his moral responsibility. He can still be considered responsible for not accompanying her, just to a lesser extent. Next, the fact that this was an isolated occurrence seems to strengthen the Coates and Swenson objection. This is because the possible world in which Leo does accompany his mother to the doctor’s office is significantly close. There would only need to be a minute change – perhaps the client having an extra available day – for him to have actually prioritized his mother, thus it seems that he appears to be more

morally responsible than he was originally. If instead this was a repeated occurrence, perhaps because Leo was struggling with mental health issues and overworked himself as a result, then perhaps it could be said that he is less morally responsible. But, as it stands, it seems as though this being an isolated incident only serves to make him more responsible. Lastly, in regard to the moral luck argument, it can be seen as irrelevant. While it is true that the only reason Leo is being morally judged is because of his mother's reaction, a factor which he cannot control, it does not seem to be the case that he can be fully excused as a result. His original action of missing his mother's doctor's appointment can still be considered morally responsible, regardless of her reaction, and regardless of whether moral judgement is passed upon him.

Response

All of that is to say that Coates and Swenson's gradable framework poses a substantial objection to my counterexample aimed at MRR. The main challenge comes in the form of Leo's responsibility shifting from nonexistent to partial as a result of the nearest possible world in which he accompanies his mother to her appointment, being rather close. However, I will still attempt to push back and perhaps argue that such a response does not address all the flaws in MRR; namely, its inability to account for the external factors that undermine the authenticity of moral responsibility. While it is true that an agent does not have control over external factors, Leo's case is a prime example that they nevertheless impact how and if an agent experiences moral judgment. Again, consider that Leo would not have been seen as morally responsible if his mother did not feel distress. Indeed one can argue and say that regardless of his mother's reaction his action was still wrong, but I do not believe that this is so clear cut. It seems that if it was not for his mother feeling distressed, Leo's moral responsibility would not have been questioned whatsoever.

Further, consider a possible world in which Leo's mother was not distressed that

he could not accompany her to her appointment. In this world, the reason for Leo to accompany her is no longer morally significant, since it does not exist. Further yet, consider a world in which Leo's mother was fully dependent on him because she could not speak English and needed him as a translator–assume that there is no one else that could go with her and that this is the only day she can go because she is about to run out of medication. Let us say that if he did not come with her she would still be distressed, and that her distress is still the reason he considers. Even then, it seems that in the world where she is dependent on him her distress seems to be a more significant reason than it was in the world where she did not depend on him. Additionally, it seems certain that if Leo were to miss this appointment where his mother was dependent on him, we would consider him more responsible than we did originally. This shift in our understanding of his responsibility is despite the fact that the possible world in which Leo (in the world where his mother is dependent on him) and Leo (in the world where his mother is independent) attend their mother's appointments remains at the same proximity. Again, in both worlds all that would need to be changed is the client's flexibility.

The reason I mention this is to attempt to show how important context is to moral judgement, and the reasons that are being considered. I am not saying this is a perfect response, instead that it could be something that a view of MRR, and Coates and Swenson's view of gradable moral responsibility, could consider.

Conclusion

While my counterexample to the sufficiency of MRR may not be without its flaws, I hope I have sufficiently demonstrated that its role in grounding moral responsibility can be challenged. To aid my argument, I first began by providing an outline of Fischer and Ravizza's account of guidance control and its constituent parts. I then narrowed my focus particularly to the reasons-responsiveness aspect of guidance control and explained what is required for a

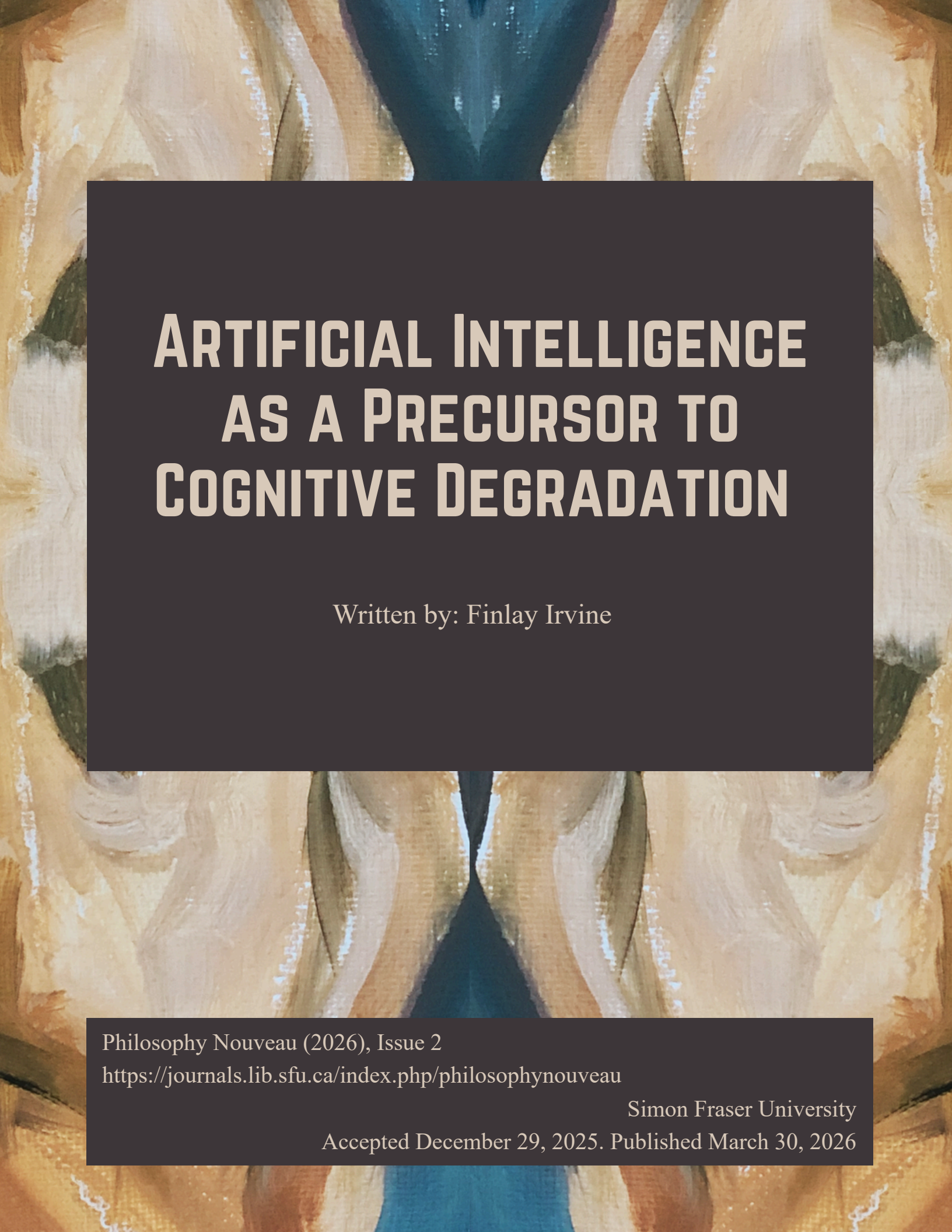
mechanism to be reasons-responsive, namely, weak reactivity and regular receptivity. I then provided a counter example in which an agent fulfilled both requirements for reasons-responsiveness, and yet did not seem to be intuitively morally responsible. Afterwards, I introduced Coates and Swenson's view of gradable moral responsibility, and explained how it could be used as an objection to my counterexample. Next, I considered a possible response to said objection, hopefully showing that MRR may have some flaws that should be considered.

References

Coates, D. Justin, and Philip Swenson. "Reasons-Responsiveness and Degrees of Responsibility." Philosophical Studies, vol. 165, no. 3, 2013, pp. 629-45.

Fischer, John Martin, and Mark Ravizza. "Chapter 3. Responsibility for Actions: Moderate Reasons-Responsiveness," Responsibility and Control: A Theory of Moral Responsibility, Cambridge University Press, 1998.

Fischer, John Martin, and Mark Ravizza. "Review of Responsibility and Control: A Theory of Moral Responsibility by John Martin Fischer and Mark Ravizza," Philosophy and Phenomenological Research, vol. 61, no. 2, Sept. 2000, pp. 441-45.



Artificial Intelligence as a Precursor to Cognitive Degradation

Written by: Finlay Irvine

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau
Simon Fraser University
Accepted December 29, 2025. Published March 30, 2026

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau

Abstract

In recent years, AI technology has proliferated, becoming unavoidably integrated into the lives of people across the globe. As this technology has grown omnipresent, a debate surrounding its merits has emerged. Most mainstream contemporary discourse on Artificial Intelligence has treated this technology as a benign, if not benevolent, development, with much of the existing literature prematurely dismissing the severity of the costs. As social forces pressure for widespread adoption of this technology, society has begun to acquiesce. Should the pervasive acceptance of AI be a concern? Drawing from influential thinkers such as Vallor, Kant, Rousseau, and Aristotle, By identifying the risks AI poses to our , this article acts as a compelling warning against the thoughtless adoption of AI technology.

Simon Fraser University
Accepted December 29, 2025. Published March 30, 2026

Artificial Intelligence as a Precursor to Cognitive Degradation
Finlay Irvine

Introduction

Progress and improvement are not universally correlated. Progress, the ability to improve, invent, and discover – to eradicate those aspects of human existence that cause suffering, struggle, turmoil, and even inconvenience – may, in reality, be the condition that causes the deterioration of innate human faculties. This is the theory purported by Rousseau, who believes that the human faculty of self-improvement results in the development of luxuries, which, by way of the temptation of their convenience, pervade society, resulting in humans developing a reliance on these luxuries that detracts from their natural faculties. This theory regarding the insidiousness of progress can effectively illuminate the potential vices associated with technological developments such as artificial intelligence (AI) that have proliferated in recent years. As AI becomes integrated into the daily lives of regular people, exempting them from both the most mundane and cognitively stimulating aspects of human existence, the impact of this now prevalent technology on human autonomy, introspection and the cognitive faculties must be considered.

Vallor on Technology and the Examined Life

The outsourcing of previously human tasks to technology has been a gradual process spanning decades. It has changed how we allocate our time, our interpersonal relations, and even the manner in which we reflect upon our lives. It is this last modification of the human experience that concerns Shannon Vallor's treatise Technology and the Virtues, within which she examines the implications of technology on human introspection and the nurturing of individual virtue. She champions the cultivation of the examined life, a process of self-reflection that aligns "one's actions, values, emotions, and beliefs with the Good," ultimately resulting in one's moral flourishing. (179) She then examines contemporary movements of self-examination that rely on technological assistance such as the quantified self movement, which uses "devices to measure, track, analyze, and

store volumes of recorded data concerning an ever-expanding list of personal variables." (199) The personal dataset produced by such technology could work in tandem with artificial intelligence, "[telling] us what adjustments we need to make to our behavior or thinking." (203)

Proponents of the quantified self movement would argue that technology improves the human capacity for virtue by instructing us on the right course of action, potentially fostering the examined life for which Vallor advocates. Vallor, however, refutes this assumption, arguing that reliance on technological assistance in matters of self-reflection is antithetical to the examined life insofar as "a dataset is not a life at all…my life includes my future, and thus the examined life is always a project, never an achievement." (203) Reliance on AI in matters of decision-making produces a moral path dependency that bypasses the process of self-reflection and constrains individual growth; according to Vallor, the key to self-examination is that it be conducted by the self rather than a detached third-party entity such as AI. Despite making a compelling case against technologically guided self-reflection, Vallor simultaneously downplays the concern that humans could become morally subservient to AI. She denies that "apps, wearable computing, or 'smart' environments make us into mindless moral zombies." (204) On her view, while AI might detract from the examined life, it will never replace human moral autonomy. However, this a priori dismissal of humanity's growing dependency on technology disregards the pernicious threats AI poses to the cognitive faculties of intellect, reason, and introspection. Ultimately, such technological reliance is markedly less inconceivable than Vallor claims.

Kant, Immaturity, and the Appeal of AI

The likelihood of such cognitive dependency is evidenced by the similarities between the enticing nature of AI and the desire to remain immature described by Kant in his essay on enlightenment, "An Answer to the Question: What Is Enlightenment." In contemporary society AI assumes the role of the guardians

denounced by the German philosopher. These guardians paternalistically guide people's actions, thus perpetuating widespread immaturity, or "the inability to make use of one's intellect without the direction of another." (17) However, while these guardians underpin immaturity, Kant believes this dependency to be "self-incurred" in that "its cause does not lie in a lack of intellect, but rather in a lack of resolve and courage to make use of one's intellect." (17) He identifies "[i]dleness and cowardice [as] the reasons why such a large segment of humankind…is nonetheless content to remain immature for life; and these are also the reasons why it is so easy for others to set themselves up as their guardians" (17). Idleness or laziness, which Kant discerns as being one of the key aspects of the human condition underpinning immaturity, instills within humans the predisposition towards the convenient rather than the onerous, even when such convenience insidiously degrades essential aspects of their humanity. AI represents the most salient of conveniences, and its ability to complete both mundane and cognitively challenging tasks with which people are burdened demonstrates its unique capacity to exploit human laziness.

Just as humans once outsourced the burden of exercising their own intellect, autonomy, and reason to guardians, they now begin subordinating themselves to AI. The burgeoning dependence on AI represents a movement antithetical to enlightenment, in which people succumb to their impatient nature and their desire for intellectual comfort rather than challenge. It is this aversion to challenge that characterizes humanity's adoption of AI and illustrates the manner in which cowardice motivates immaturity. When the exercise of one's own intellect is perceived as a burden rather than an integral aspect of the human experience, our tendency towards cowardice prompts us to exchange enlightenment for self-incurred immaturity. Herein lies the challenge to enlightenment, namely in that "it is so comfortable to be immature." (17) To progress towards maturity, or independent thought, would require overcoming what has become viewed as an ostensibly insurmountable challenge; throwing aside the shackles of cognitive deference and entering the unknown of intellectual sovereignty. The potential

discomfort associated with this ordeal acts as a deterrent that prevents humankind from achieving maturity. Considering AI promises to more effectively alleviate the onus of decision-making and introspection, the human tendency towards immaturity, and the discomfort associated with relinquishing this immaturity, will only be exacerbated.

Which Forms of AI Reliance Are Most Dangerous?

Despite humanity's escalating reliance on AI, its potential consequences have not been clearly identified. It is first essential to differentiate between the types of activities AI could perform and the corresponding consequences of each. The outsourcing of menial activities to AI, such as those done thoughtlessly by humans and often accompanied by cognitive distractions such as music or background entertainment, pose a merely superficial threat to human flourishing insofar as such activities are not vital to the preservation of integral human faculties. Not only do these activities require little cognitive stimulation, their tedious nature might also exhaust reservoirs of cognitive energy. While AI's ability to alleviate humanity of such tasks may still pose inherent risks, efforts should be reserved for resisting the adoption of AI with respect to more paramount human capacities. These include the reliance on AI for activities of significant intellectual exertion, such as writing or artistic ventures, and for matters of decision-making and self-reflection that could be alleviated by AI life coaching. Such reliance could ultimately undermine the cognitive capacity for reason, introspection, autonomy, and virtue, that are vital aspects of the human experience. Moreover, considering the human propensity towards self-incurred immaturity, which AI technology will only accentuate, several potential consequences on the cognitive faculties must be scrutinized.

Aristotle on Virtue, Reflection, and Self-Sufficient Thought

The first such consequence is that the reliance on AI precludes the cultivation of

virtue and is, of itself, a vice. In The Nicomachean Ethics, Aristotle argues that happiness, the ultimate end, is achieved by the "exercise of the vital faculties in accordance with excellence or virtue" (Aristotle, 32). Achieving virtue, therefore, is a necessary precondition to living a happy, prosperous life. Virtue is cultivated by emulating acts that correspond with the desired virtue. Aristotle defines the virtuous act to be "a kind of moderation, in as much as it aims at the mean," (57) while there exist two diametrically opposed classes of vices, "one marked by excess, the other by deficiency." (63) However, while both vices are opposed to the mean, they are not equally so. Aristotle explains that "those things to which we happen to be more prone by nature appear to be more opposed to the mean," and that insofar as "one extreme is, in fact, nearer and more similar to the mean, we naturally do not oppose it to the mean so strongly as the other."(64) Determining which vice one has a natural predisposition towards requires introspection and reason, the sort incompatible with a reliance on AI, insofar as introspection must necessarily be conducted by oneself. A reliance on AI in matters of self-reflection renders the individual incapable of aligning themselves with the mean and achieving virtue in all aspects of their life.

Not only does AI impede the introspection required to discern and assess one's own vices, it itself represents a vice that must be opposed. In the sphere of independence of thought, reason and self-sufficiency represent a clear virtue, inasmuch as both appear to be prima facie prerequisites to happiness: one must have a degree of independence and free thought to exert mastery over the course of their life. The vices to this virtuous mean are immaturity on the one hand, and obstinacy on the other. Therefore, what is integral to self-sufficient thought is that one is neither intellectually reliant on others nor resistant to dialectical discourse. The use of AI as a decision-making tool, which necessitates the relinquishing of self-sufficient thought to another entity, aligns with the vice of deficiency. Insofar as it has been proven, by analyzing Kant's call for enlightenment, that humans have a natural tendency towards self-incurred immaturity, it is evident that this vice more starkly opposes the mean of self-sufficient reason. Therefore, not only is the

immaturity associated with a reliance on AI a vice in and of itself, but it is also the vice to which one must most ardently oppose in order to align oneself with the virtuous act, which, via its consistent exercise, precipitates virtue and happiness.

In addition to requiring that one consistently practice the virtuous act, Aristotle outlines a second precondition to achieving virtue, namely that the individual must "be in a certain state of mind when he does [the act]." (51) To achieve this state of mind the individual must first "know what he is doing; secondly, he must choose it, and choose it for itself; and thirdly, his act must be the expression of a formed and stable character." (51) Proponents of AI might endorse its ability to instruct an individual on the right, or virtuous, course of action. However, a reliance on AI to determine which course of action is correct necessarily precludes the formation of the state of mind requisite to fostering virtue. Someone who outsources the process of deducing the virtuous act can neither claim to know 'what he is doing' nor why he is compelled to act that way. Moreover, this person is clearly not choosing the act for itself insofar as they are merely following the directions of AI rather than any internal belief in the goodness of that act. Finally, subservience to AI instruction is incompatible with the formed and stable character Aristotle describes, namely in that it is impossible for a person to internalize the virtue of an act if they have not engaged in the introspection required to determine why that act is virtuous in the first place. Ultimately, regardless of its ability to deduce the virtuous act, the dependency on AI impedes the cultivation of virtue because it allows one to bypass the process of reflection requisite to fostering the virtuous state of mind.

Rousseau on Progress, Luxury, and Human Decline

Likewise to Aristotle, Jean-Jacques Rousseau's The Social Contract and Discourses offers valuable insights pertaining to the risks associated with the emergence of AI technologies. In his essay Rousseau describes the insidious evolution of man from the state of nature to the state of society, and the corresponding process of

degradation he endures. Within the state of nature man's faculties are highly attuned, particularly those physical ones conducive to survival, and he is incapable of reason. He also possesses the faculty of self improvement "which, by the help of circumstances, gradually develops all the rest of our faculties" including reason. (169) While this faculty might appear to be beneficial, Rousseau, however, believes it to be "the source of all human misfortunes; that it is this which, in time, draws man out of his original state, in which he would have spent his days insensibly in peace and innocence." (169) It is this faculty that induces the establishment of society, and with it the luxuries and conveniences that so degrade the natural human form. These new advances and conveniences over time lose "all their power to please, and even degenerate into real needs…enervat[ing] both body and mind," (186) and ultimately engendering "the decrepitude of the [human] species." (187)

AI then, on this view, represents only the most recent in a long line of advancements that corrupt human faculties. Just as the ladder reduced the human capacity for climbing, or the calculator inhibited mathematical ability, so too will AI deteriorate the human faculties of cognition, most saliently the capacity for reason. However, Rousseau would contend that the consequences of AI on the cognitive faculties should not be a preeminent concern. Instead, he argues that insofar as the "state of reflection is a state contrary to nature… [the] thinking man is [therefore] a depraved animal."(167) To Rousseau, man in the state of nature is a purely instinctual being, one whose actions are dictated not by thought, but instead by natural impulse. Modern man, by comparison, has developed, by way of his faculty of self improvement, the capacity for contemplation. However, instead of engendering the prosperity of man, Rousseau believes that by thinking, man withdraws himself from the state of nature and inflicts upon himself "fatigue, mental exhaustion, [and] innumerable pains and anxieties." (167) In brief, Rousseau views the development of the cognitive faculties as the culprit for modern man's incurable suffering, and insofar as AI is capable of revoking these faculties, he would endorse the technology as a gift rather than a pernicious threat.

Why Reason and Autonomy Must Be Preserved

Rousseau's denouncement of the cognitive faculties would be ardently disputed by both Aristotle and Kant, who believe that thought, reason, and introspection are integral aspects of our humanity and therefore should not be forfeited to AI, regardless of their presumed absence in the state of nature. Aristotle, for his part, holds the contemplative life as the life most conducive to happiness, which he views as the ultimate end. He argues contemplation is uniquely self-sufficing in that one "is able to speculate by himself" whereas virtuous actions need people towards whom they can be directed. (343) Additionally, because "happiness is thought to imply leisure" and virtuous actions are restless while contemplation is not, therefore "happiness…consists in [contemplation and] the exercise of reason." (343) Considering its function as a prerequisite to happiness, it is, therefore, imperative that the human capacities of reason and contemplation be sustained, and the concern that AI subordination might contravene such faculties be heeded.

Kant similarly believed the faculties of reason and intellect to be essential to human flourishing. Underpinning his call for enlightenment is the idea that respecting the free thought of the individual is necessary to "treat[ing] the human being…in accordance with his dignity." (23) By depriving the individual of free thought, which includes the autonomous use of their reason, the dependency on AI contravenes the dignity inherent to every human being and thus poses a profound and unique threat. Moreover, Kant posits that free thought acts as a necessary precursor to freedom of action. By enfeebling intellectual autonomy, AI simultaneously threatens the total sovereignty the individual holds over themselves. Regardless of the unnaturalness of those faculties by which we form autonomous thought, reason, and introspection, they have become vital components of human nature that enable us to act independently, and achieve moral virtue, happiness, and dignity. By virtue of the prosperity they enable, it is pivotal that the threat posed by AI towards them not be dismissed.

Conclusion

In “An Answer to the Question: What Is Enlightenment,” Kant raises another salient concern, namely that it is especially “difficult for any individual to work himself out of the immaturity that has almost become second nature to him.” (18) Even if this current generation is capable of resisting the impulse to adopt AI and relinquish the burden of free will and reason, the following generations who grow up with this pernicious technology as a celebrated, aspect of their society might become incapable of enlightenment insofar as technological reliance has become ingrained in their understanding of what it means to be human. Should such a dystopian nightmare become a reality, then human beings will be relegated to the passenger seat of their own existence, while AI dictates the course of their lives, their human autonomy revoked, and their intellectual self-sufficiency degraded beyond recovery. In such a case the human experience would be degraded beyond the point of recognizability, echoing the solemn lessons of Rousseau. Even if such a catastrophe is avoided, it becomes clear from the warnings of Kant, Rousseau, and Aristotle that we must reevaluate our relationship to AI lest we risk the degradation of those faculties that make us human.

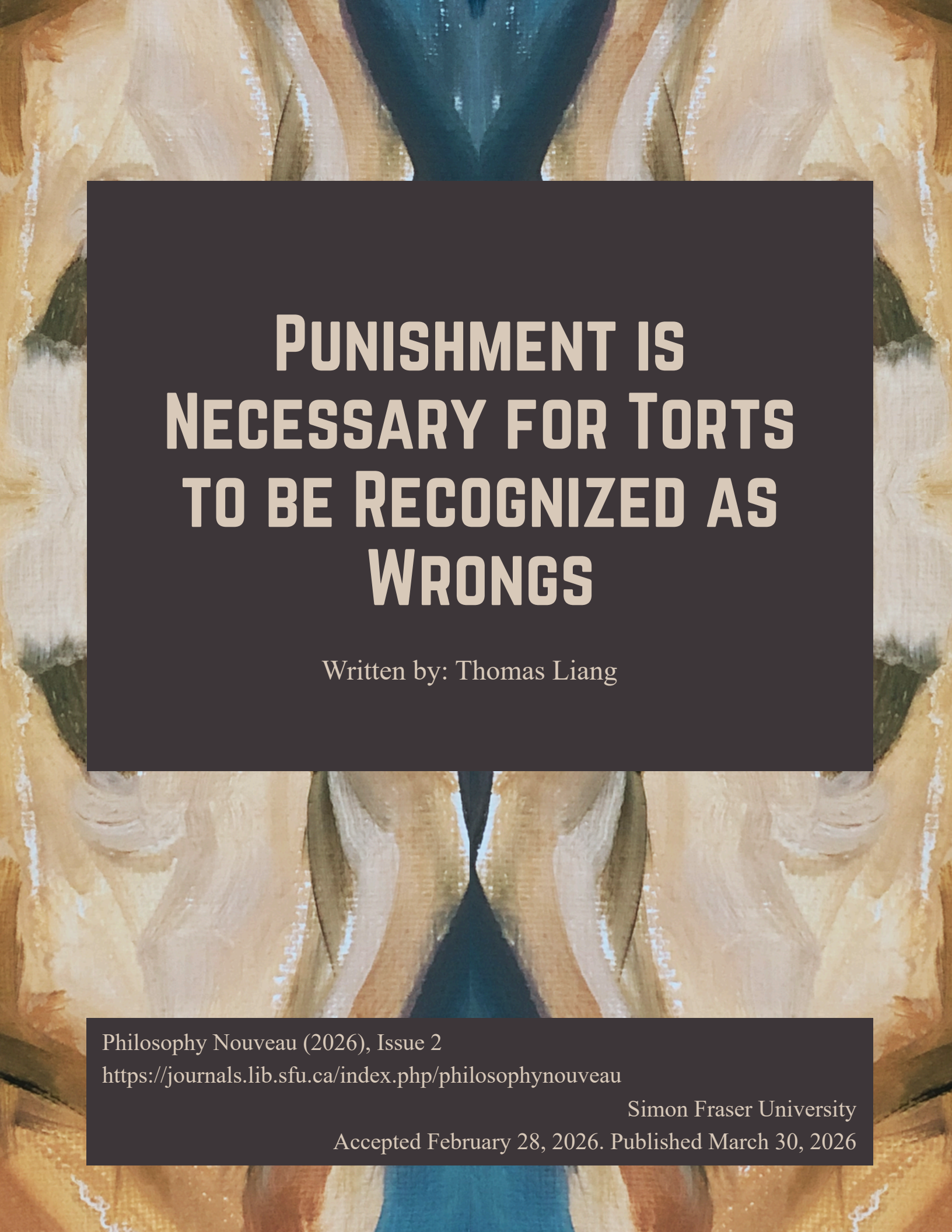
References

Aristotle. The Nicomachean Ethics. Translated by F.H. Peters. London: Kegan Paul, Trench, Truebner & Co., 1893.

Kant, Immanuel. “An Answer to the Question: What Is Enlightenment.” Essay. In Toward Perpetual Peace and Other Writings on Politics, Peace, and History, 17–23. New Haven, Connecticut: Yale University Press, 2006.

Rousseau, Jean-Jacques. The Social Contract and Discourses. Translated by G.D.H. Cole. New York: J.M. Dent & Sons, 1923.

Vallor, Shannon. Technology and the Virtues. Oxford: Oxford University Press, 2016._



Punishment is Necessary for Torts to be Recognized as Wrongs

Written by: Thomas Liang

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau
Simon Fraser University
Accepted February 28, 2026. Published March 30, 2026

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau

Abstract

In this essay, I argue that Goldberg and Zipursky's resolution of the Inaptness-of-Liability Problem (ILP) for wrongs-based theories of tort is inadequate. The ILP holds that torts should not be understood as wrongs, but rather, as losses. If they truly were wrongs, we would expect our legal system to respond with criminal punishment rather than the civil remedies we actually see. I begin an argument for the inadequacy of this view by outlining the problem as the authors present it. I then discuss their proposed solution: that punishment is not a necessary component of holding a wrongdoer accountable. I contend that, because Goldberg and Zipursky do not sufficiently justify this claim, their response fails to neutralize the problem. Thus, the ILP remains compelling.

Simon Fraser University
Accepted: February 28, 2026. Published: March 30, 2026

Punishment is Necessary for Torts to be Recognized as Wrongs
Thomas Liang

Introduction

In this essay, I argue that Goldberg and Zipursky's resolution of the Inaptness-of-Liability Problem (ILP) for wrongs-based theories of tort in the 2010 publication "Torts as Wrongs" is inadequate. Drawing on the work of Jules Coleman and Stephen Perry, the ILP holds that torts should not be understood as wrongs. If they truly were, we would expect our legal system to respond with criminal punishment rather than the civil remedies we actually see. In practice, tort law reallocates losses by shifting the cost of the injury from the plaintiff back to the defendant. Because punishment is absent and the civil remedy arises from a "duty of repair" that addresses losses rather than wrongs, torts appear more accurately characterized as losses than as wrongs. (Perry 8) I begin an argument for the inadequacy of this view by outlining the problem as the authors present it. I then discuss their proposed solution. Finally, I contend that, because Goldberg and Zipursky do not sufficiently justify why punishment should not be an integral component of tort law, their response fails to fully neutralize the problem. Thus, the ILP remains compelling.

Goldberg and Zipursky's Response to the Inaptness-of-Liability Problem

Goldberg and Zipursky think that the conclusion drawn by the ILP is "unwarranted." (14) They maintain that tortious wrongs and criminal wrongs are distinct types of legal wrongs, and as such, the legal system is justified in addressing them differently. In particular, tortious wrongs are distinctive because they are relational, private wrongs: they create normative connections between individuals themselves. When a tort occurs, there is a tortfeasor and a wronged party; the plaintiff and defendant are thereby situated in a specific normative relationship. The state, in this situation, acts as a neutral arbiter. This, Goldberg and Zipursky argue, distinguishes torts from criminal or regulatory wrongs, which are primarily public wrongs against the state or society at large. Crimes, in this sense, are not directed against any particular individual.

Against this backdrop, the authors contend that it is not difficult to justify why torts warrant a different kind of legal response than crimes. On their view, it makes sense that state prosecution is different from a plaintiff exacting money damages from a defendant when he successfully sues. Goldberg and Zipursky agree with Coleman and Perry that wrongdoers should be held responsible for their wrongdoing – however, their key point is that both civil liability and criminal punishment are different forms of accountability employed by our legal system. Before proceeding with my critique, I wish to first clarify Goldberg and Zipursky's view on the relationship between a wrong and punishment. They maintain that although punishment is "in some sense about responding to wrongs," there is no necessary connection between the two.
(8)That is, torts can be realized as wrongs without involving any element of punishment, contra Coleman. The authors do not make this claim explicitly, but it can be inferred from their presentation of civil liability and punishment as two separate yet valid mechanisms by which the legal system holds wrongdoers accountable.

Accountability and the Limits of Compensation

My critique, then, is that Goldberg and Zipursky do not adequately explain why punishment need not be connected to the rectification of a wrong when holding a wrongdoer accountable. Although this claim functions as a core assumption underlying their argument, the authors neither explicitly acknowledge it nor offer a clear justification for why punishment is not a necessary component of rectifying a wrong. I share the opinion that wrongdoers in both tort and criminal cases ought to be held accountable for their wrongdoing. But what does it mean to be held accountable for a wrong? By exploring this question, I seek to illuminate some aspects of the issue that Goldberg and Zipursky do not fully address and to elucidate how my rationale departs from theirs.

Suppose that torts are to be recognized as wrongs. That means that every time I

commit a tort, I am committing a wrong. When I commit a wrong, then, there is a widely accepted belief that I should be held accountable for my actions. In my view, this criterion of "accountability" for a wrong ought to be satisfied by two conditions. First, (A1): after committing a wrong, I should be required to make reparations to the injured party. In tort law, this typically means employing the default remedy of "making the plaintiff whole" via compensatory damages. This restores the plaintiff to the position they occupied before the wrongful loss, while divesting me of any wrongful gain. This is the first condition for accountability, and as Goldberg and Zipursky suggest, this is all that is required for holding tortfeasors accountable. As noted, however, they do not provide an explicit explanation for why this should be the case, and I personally do not find it intuitively compelling.

If torts are to be understood as wrongs, I argue that the first condition alone is insufficient to satisfy the criterion of accountability. Like Coleman, I hold that compensating the plaintiff and depriving the defendant of any wrongful gains merely restores the parties to their original, uninjured positions—this is what the first condition achieves. While this process of "making whole" amends the relations between the plaintiff and defendant, it does not, in my view, adequately communicate the stronger message that the defendant's conduct was wrong.

Why Wrongs Require Punishment

For this reason, I believe a second condition may be necessary. If what the defendant, did was wrong, then I should be reminded by the state or by society otherwise that the behaviour I exhibited is not socially acceptable nor encouraged. Thus, I posit (A2): the state should inform me, on the plaintiff's behalf, that this behaviour is socially inappropriate and should be deterred to prevent similar future wrongs. How else should I be reminded or this conduct discouraged? Punishment seems the natural answer. Part of holding a wrongdoer accountable (as opposed to merely a loss-doer, for example; I am again relying here on our common-sense understanding of a "wrong" versus a "loss") requires conveying the

societal message that such behaviour will not be socially tolerated. Accordingly, I contend that an additional legal function – namely, punishment – is necessary for torts to be recognized as wrongs.

Objection and Reply

A possible objection from a defender of Goldberg and Zipursky's view is as follows: "But the authors do provide an (incidental) justification for the first condition! They contend that the purpose of tort law is to provide an avenue for citizens to seek civil recourse after being wronged. Because punishment is plausibly not a major feature of civil recourse – appearing above and beyond compensation – this justifies the use of only the first condition as necessary for accountability in tort law."

This objection simply collapses into the broader "accountability" issue discussed above. Recall that my challenge to the authors was this: If torts are to be understood as wrongs, and not losses, then one must provide justification for excluding punishment from the process of rectifying a wrong. A response such as, "The purpose of tort law is to provide citizens with an opportunity for civil recourse, and punishment is plainly not part of civil recourse," seems to be a non sequitur. It sidesteps and fails to defend the very claim that Goldberg and Zipursky themselves advance: that torts should be recognized as wrongs rather than losses. Indeed, if my preferred "torts-as-losses" view were adopted, I would readily accept the objection above, since, in my view, punishment is not a necessary component of holding someone accountable for a loss. My concern with the authors' omission of punishment arises only because they insist that torts are wrongs. But if torts were instead conceived purely as losses, the exclusion of punishment would be justified. Moreover, the objection unravels upon closer inspection. What does it mean, for instance, for citizens to seek recourse? The objector might say that it involves a right to be redressed for one's grievances and for the perpetrator to be held accountable in court. But once civil recourse is understood to include some notion

of the wrongdoer's accountability, we are led straight back to the central question explored in this essay: What counts as accountability?

Conclusion

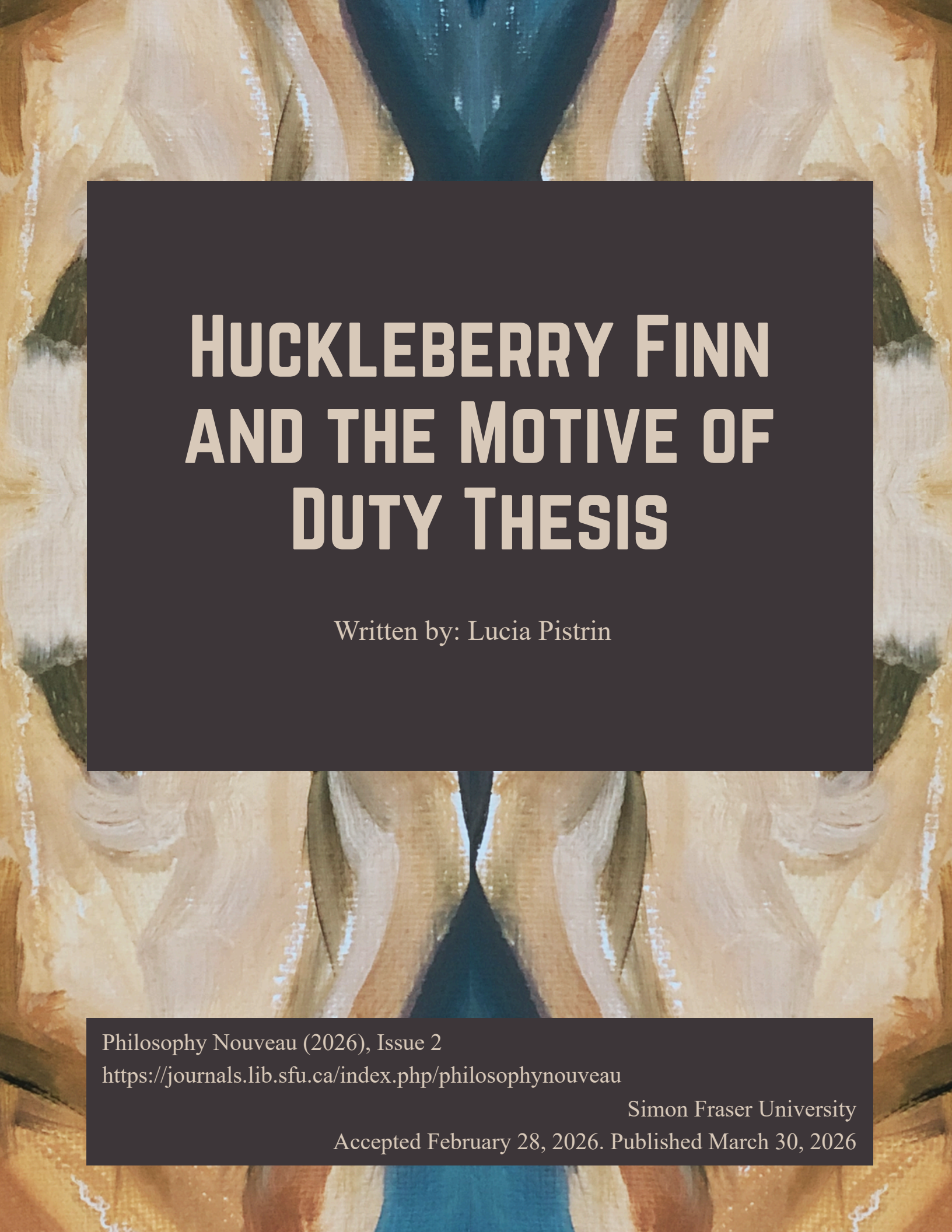
In conclusion, Goldberg and Zipursky contend that torts should be understood as wrongs rather than losses, and they defend this view by claiming that punishment is not a necessary component of holding a wrongdoer accountable. I have argued to the contrary. Simply reallocating losses may repair the relations between the parties, but it does not sufficiently communicate that the tortious conduct was socially unacceptable or that such conduct ought to be deterred in the future. This is one such difference between understanding torts as wrongs versus torts as losses. Finally, I have considered and addressed a potential objection, arguing that it fails to engage the central issue and ultimately collapses into circular reasoning.

References

Coleman, J. (1992). Risks and Wrongs. Cambridge University Press.

Goldberg, J., & Zipursky, B. (2010). Torts as Wrongs, vol.88. https://ir.lawnet.fordham.edu/faculty_scholarship/673

Perry, S. (1992) The Moral Foundations of Tort Law. https://scholarship.law.upenn.edu/faculty_scholarship/1153



Huckleberry Finn and the Motive of Duty Thesis

Written by: Lucia Pistrin

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau
Simon Fraser University
Accepted February 28, 2026. Published March 30, 2026

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau

Abstract

In “Acting for the Right Reasons,” Julia Markovitz objects to the Kantian Motive of Duty Thesis (MDT) and offers her own Coincident Reasons Thesis (CRT) in its stead. The MDT states that for an action to be morally worthy it must be performed from duty to the moral law. On this view, the MDT would fail to attribute moral worth to Huckleberry Finn’s actions in saving Jim from servitude to Mrs. Watson. I turn to Kant’s original work on the subject in *The Groundwork* to argue that this is not the case. When the integral connection between the Motive of Duty Thesis and the principle of humanity is spelled out it can be made explicit that Huck acts from duty when he rescues Jim from servitude to Mrs. Watson – thus, Finn’s actions can rightly claim moral worth

Simon Fraser University
Accepted: February 28, 2026. Published: March 30, 2026

Huckleberry Finn and the Motive of Duty Thesis
Lucia Pistrin

Introduction

In this paper, I will outline the reasons for which I believe that the Huck Finn case does not present an issue for the Motive of Duty Thesis (MDT). The Motive of Duty Thesis was first introduced by Immanuel Kant in The Groundwork of the Metaphysics of Morals, who maintains that an act only has "genuine moral worth" when it is performed "without any inclination, simply from duty." [398] I suggest here that Huck's actions are not at odds with Kant's view respecting acting from duty, and stem from an implicit commitment to the principle of humanity. When the integral connection between these concepts is spelled out, it is clear that Huck acts from duty when he rescues Jim from servitude to Mrs.Watson. Thus, his actions have moral worth.

In "Acting for the Right Reasons," Julia Markovitz uses the Huck Finn example to motivate an objection to the traditional Kantian Motive of Duty Thesis and proffer her own Coincident Reasons Thesis (CRT) in its stead. The MDT as reconstructed by Markovitz claims that for an action to be considered morally worthy that action must be performed from a duty to the moral law. According to the CRT, an agent's right action has moral worth just in case the "motivating reasons for acting coincide with the reasons morally justifying the action." (205) The distinction between the MDT and the CRT makes sense of Markovitz' suggestion that a morally attractive person will help others not because they ought to obey the moral law but because those others are in need of help, a claim that feels intuitive. (204) This aligns with the Right Reasons Thesis (RRT) which states that "morally worthy actions must be performed for the right (motivating) reasons." (203)

Why Huck Finn Seems to Challenge the Motive of Duty Thesis

According to Markovitz' view, the Motive of Duty Thesis would problematically exclude the Huck Finn case by failing to attribute moral worth to his action. This is

because it was not performed from a duty proper to perform the right action; in fact, it was performed from significant and persistent unease and a conviction that he was doing the wrong thing. This act is a form of inverse akrasia. As Nomy Arpaly suggests, these sorts of acts occur when an agent does the correct thing despite their better judgment urging them to do otherwise. (75) Markovitz argument for the insufficiency of the CRT in accommodating Huck Finn's case works as follows; Huck does not know that he is acting rightly, indeed, according to the societal conventions at the time, he is acting quite wrongly. He fears eternal damnation as the inevitable consequence of this misdeed. (Twain XVI) He is not aware of the moral law largely abided by today, which bars individuals from using others as slaves or treating them differently based upon the colour of their skin. Thus, because he does not know that he has a moral duty to rescue people from slavery or to treat people of colour as equals, he cannot be said to be acting from that duty, so his action has no positive moral worth.

Markovits maintains that the reason that rescuing Jim from his slave-owner was correct was the same reason that Huck acted for - that Jim has value as a human being - and thus can rightly be considered morally praiseworthy under the CRT but not the MDT. Contra Markovitz, I am of the mind that if the basis on which Huck acted was the reason that Jim has value as a human being, his actions are aligned with the MDT, even if he was unaware that he was acting from any duty. Thus, the CRT need not be posited to compensate for this alleged shortcoming of the MDT. Huck values Jim as a human being. There is plenty of textual evidence in The Adventures of Huckleberry Finn that exhibits Huck's regard for Miss Watson's slave. This regard extends beyond a consideration of Jim as valuable qua slave, or valuable only for the services he provides. Examples of Huck's feelings that indicate an overall commitment to Jim occur in Chapter XVI, after he has performed the initial rescue by boat. Huck's inner conflict begins with a sudden concern that he has done the wrong thing and resolves in the realization that Jim's value trumps these initial, societally ingrained doubts. The scene in which Huck is approached by men on another skiff is a turning point that makes concrete the triumph of this

moral realization.

Huck begins his journey to conscious realization of Jim's value via a sudden, intense doubt to the contrary. Notably, this piece of reasoning occurs after his operation to rescue Jim is already well underway. His conscience puts his alleged wrongdoing to him clearly, telling him: "But you knowed he was running for his freedom, and you could a paddled ashore and told somebody." (Twain XVI) Further, and more personally, Huck is overcome by guilt for robbing Mrs. Watson of her property: "What did that poor old woman do to you that you could treat her so mean?" (Twain XVI) This is evidence of a temporary conviction that rescuing Jim was gravely wrong. Huck commits to turning Jim in at the first opportunity to relieve himself of this grief. He describes himself as feeling "good and all washed clean of sin for the first time… in my life." (Twain XVI) But when the opportune moment faces him he is unable to act in the way that he has resolved to do. Feeling his motivation weakened, he lies and tells the men that his father is underneath the boat with small pox, and Huck and Jim sail away scot-free.

Huck's Real Motivation: Jim's Value as a Human Being

What I wish to stress here is that Huck's belief that his action was wrong was only temporarily motivating. It motivated him to consider, very strongly, the option of turning Jim in. But when push came to shove, Huck was not swayed enough by this operation of his conscience to reverse the course of his action and betray his friend. Therefore, contrary to Nomy Arpaly's suggestion, Huck's action should not be considered an instance of inverse akrasia. It was not weakness of will that made Huck refrain from turning in Jim. Instead, it was a stronger motivating factor, namely, a recognition of Jim's value. Huck internally recounts Jim's goodness to himself in an excerpt from Chapter XVI after he has resolved to write to Miss Watson. In his mind's eye, he sees "Jim before me all the time… talking and singing and laughing… always call me honey, and pet me and do everything he could think of for me, and how good he always was." (Twain XVI) Behind this broken grammar is

a concrete consideration of Jim as valuable qua friend. This commitment extends beyond a mere whim, as Huck claims that if the situation were repeated he would have acted in the same way again. By acting to save Jim, Huck does not simply fail to turn Jim in; instead, his action is morally praiseworthy because he realizes how Jim's value trumps the societal norms put forward to him by his conscience. This is evidenced by the conscious effort Huck puts forward in concocting an elaborate tale to deceive the men and remaining unwavering in interrogation. This act embodies serious grit, resolve, and commitment to his fellow man.

In "The Conscience of Huckleberry Finn," Jonathan Bennett construes Huck's motivation and the sways of his conscience as exemplifying the relationship between 'bad morality' and 'sympathy.' A bad morality, for Bennett, is a set of precepts that an agent adheres to that differ from the author's. He neglects to posit an objective moral standard, but rests on the assumption that the reader would agree with his disapproval of these sorts of moral codes. An example of a bad moral code that he offers is the mentality of slave-owners in the deep South. Bad morality is contrasted with 'sympathy,' to which Bennett refers to visceral feelings that impel action from an agent, most often as a result of witnessing somebody else suffering. Bennett describes scenarios where one's sympathies have the potential to impede clear-minded moral judgments, as in the case of a mother forced by an understanding of her small child's condition to hand him over to a doctor despite his tortured wailing at the prospect. (Bennett 124) If she were to turn herself over to her feeling of sympathy for her child, she might be tempted to keep him away from the doctor to mitigate his short-term suffering. The distinguishing feature of this thought experiment from the textual and historical examples Bennett draws from in the bulk of the work is that the mother is here conflicted between a good morality and sympathy, rather than a bad one.

Bennett maintains that in the passage examined above in Chapter XVI of the Adventures of Huckleberry Finn, Huck is caught in a cross-fire between bad morality and sympathy. The bad morality in question is the Southern mindset

towards people of colour, ingrained in Finn's consciousness since youth. The sympathy he feels is made apparent in the scenarios that run through his mind at the opportune moment, which compel Huck to continue acting to save Jim despite his temporary doubts. Bennett cites the following passages from Chapter XVI as evidence for his claim that in Huck's case, sympathy wins over a bad morality: "Pooty soon I'll be a-shout'n for joy, en I'll say, it's all on accounts o' Huck I's a free man," and "'Dah you goes, de ole true Huck; de on'y white genlman dat ever kep' his promise to ole Jim." (Twain XVI) Jim says these things just as Huck is about to turn the former slave in. Under Benett's reading, it is these pieces of dialogue that evoke enough guilt within Huck to sway him from his temporary resolve to turn Jim in. Thus, sympathy wins out over a bad morality.

From Humanity to Duty: Why Huck Satisfies the Kantian Standard

Bennett and I differ importantly along this axis. While Bennett fails to attribute moral worth to Huck's actions by invoking sympathy, a non-moral and sometimes even counter-moral consideration, as explanation for the way Huck acts, I believe Huck is acting from something much stronger: a duty to Kant's principle of humanity. By reasoning and acting in the way that he does, Huck operates from this Kantian notion, whether or not he has any cognitive access to the doctrine Kant espouses. The principle of humanity insists that people are treated as valuable ends in themselves and are not used for merely instrumental purposes – in The Groundwork, he proffers the following: "Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means, but always at the same time as an end" as a second iteration of the categorical imperative. [429] Jim's slave owner, Miss Watson, was using him for an instrumental purpose by treating him as her slave, and thus treating him as a mere means. So despite his competing worry, the overriding concern that Huck acts upon is a duty to Jim's humanity. In doing so, he upholds both formulations of the categorical imperative test; if he passes one, he necessarily passes the other. Thus, his action is both rational and moral, because by rescuing Jim Huck both (1) operates "in such

a way that… (he could) also will that my maxim should become a universal law" [402] and (2) treats Jim as an end.

The categorical imperative is integrally related to another Kantian concept, namely, the good will. In the first section of The Groundwork, Kant states that "nothing in the world, or even beyond it, can possibly be conceived and be called good without qualification other than a good will." [393] The good will is an intention, inclination or volition towards the good. An action, for Kant, can be construed as morally praiseworthy only insofar as it is done from the good will. The good will is directed at the moral demands of duty, although the finer details of this point are a matter for live debate in Kant scholarship. Hence, the Motive of Duty thesis remains a controversial one. Kant does state that "Deviating from the principle of duty is beyond all doubt wicked." [403]

Kant crucially distinguishes between acting from duty and acting in accordance with duty. This distinction proves relevant to the assessment of moral worth. Only those actions that are done from duty rather than merely agreeing with duty or being in accordance with duty are morally praiseworthy. Kant offers an intuitive example of this distinction in The Groundwork. A prudent shopkeeper maintains fair prices so as to promote his success by attracting naïve customers. [397] Despite the fairness of his policy, he does not act from a concern for duty, and thus lacks the good will requisite for a morally worthy act. The distinction between right actions and what is morally praiseworthy is also raised by Nomy Arpaly, and plays an important role in interpretations of Kantian ethics and moral theory more broadly. Arpaly begins her discussion of positive moral worth, or moral praiseworthiness, by observing that "sometimes a person does the right thing, but we are not particularly impressed." (67) She outlines the case of Flaubert's Madame Bovary, who takes a sudden interest in performing charitable actions, not out of a desire to do the right thing per se, but from a desire to lend her public image a saint-like quality. It is for this reason that the reader is prompted to withhold the level of moral praiseworthiness the protagonist might desire. Her

charitable acts do not have the same moral worth as those same actions would if they were done for other, less self-involved reasons. As with the prudent shopkeeper, the reasons for her good acts are not the same as those that make her actions good, and thus Bovary does not act in alignment with the RRT. Accordingly, she fails both the CRT and the MDT on Markovitz' framework.

As I have demonstrated with reference to the text and strive to make explicit here, it is overwhelmingly clear that Huck is acting from a duty to Jim's value as a human being, rather than in agreement with it, when he decides against turning his friend in. If Huck was to act merely in agreement with duty, his reasoning would not have taken the course it did in Chapter XVI. In Huck's case, acting in agreement with duty could have looked like being motivated to hide the fact that he was stowing a black man aboard ship to save his image. The distinction between the reasons for which Huck actually acts versus his actions in this counterfactual scenario run parallel to the distinction Kant draws between the prudent shopkeeper and the morally good shopkeepers in The Groundwork. If he was motivated for the same reasons that the prudent shopkeeper was in maintaining fair prices, Huck would have done the right thing by accident. Doing the right thing for the simple reason that Jim has value as a friend would not have been forefront in his mind, nor would it have shifted his course of action. As Arpaly suggests, we might still be happy that Huck has saved Jim, but it doesn't seem that he deserves moral credit for his action. (71)

The Duty-to-Act-from-Duty Problem

Recall Markovitz' distinction between the MDT and the CRT. One of the key differences between the Motive of Duty thesis and the Coincident Reasons Thesis is that the MDT excludes those cases of inverse akrasia that the CRT might support. Scholars disagree as to whether or not Kant actually implicates a duty to act from duty or the moral law in his work. W. D Ross raises an important conceptual issue: the duty to act from duty might resolve into an infinite regress and a fundamental

contradiction. Robert Pippin and Henry Allison find a different kind of vicious cycle inherent in the duty to act from duty, insofar as an agent might then have a duty to act from the duty to act from duty, and so on ad infinitum. In order to resolve these glaring and related issues with one of Kant's most significant ethical claims, Michael Walschots suggests that an action performed from the motive of duty can be considered good but is certainly not incumbent on every agent. In other words, we do not have a duty to act from duty.

On Walschots' reading, there is a general obligation to virtue, but no duty proper to act from virtue. The shift from duty to obligation renders what is morally required in a more broad and forgiving light. It is thus not required that one acts from duty but instead that one strives to "acquire the moral disposition" towards moral perfection. (Walschots 60) So, in very loose terms, individuals are obligated to strive for conformity with the moral law. This is not their duty. Walschots suggests that for Kant, one only has a duty to perform those specific actions that are in conformity with duty, but not from a sense or a 'duty' to duty. Thus, the infinite regress is stopped short. As it applies to the Huckleberry Finn case, Huck's action would be considered morally praiseworthy on Walschots' reading because he has cultivated, in himself, an inward faculty for a better moral disposition than the one that guided him earlier in the narrative. At the beginning of The Adventures, the protagonist exhibits an attitude towards slaves that is more concerned with their value as means rather than ends. Huck exudes other, more general, signs of an underdeveloped moral consciousness. In chapter I, Huck tells Miss Watson that he'd rather go to hell than reside where he was with her and Widow Douglas, despite their compassionate aims in taking him in. His naivety and generally misinformed outlook on life seems to evolve over the course of the narrative to become more mature, reflective, and self-aware. Because of the upward trajectory of his inward disposition, and the resolution of that disposition in actions that facilitate and preserve Jim's freedom, Huck has achieved a softer variant of the Motive of Duty Thesis.

Objections: Self-Interest and Psychological Egoism

I will now turn to defending Huck's actions as meeting the harsher criteria for the Motive of Duty Thesis that Markovitz presents in "Acting for the Right Reasons" against some possible objections. So far, much of my discussion has gone to show that Huck acts from a duty to Jim's value as a human in his capacity as a friend, rather than in his capacity as a slave. Markovitz suggests in her paper that "Morally worthy actions are ones that reflect well on the moral character of the person who performs them," and "when we do the right thing because it happens to suit us… our action has no moral worth." (203) It might be possible to levy a criticism of Huck's actions based on the relief it brings him in either resolving to perform action A or actually performing opposite action B. In this case, action A was turning Jim in, and action B was tricking the two men on the skiff into believing that he was not stowing a slave in his boat. One could say that Huck merely acts to assuage his conscience and is simply pulled every which way by his whims and intuitions. If each of his whims and intuitions are self-gratifying in nature as the psychological egoist has it, then Huck has no moral worth at all.

A close reading of the relevant text as well as an assessment of a conceptual problem baked into the psychological egoist agenda disproves this objection. Firstly, it would be absurd to think that Huck's self-interested desires were the ultimate decision-maker in this scenario, although they did figure into a temporary goal of turning Jim in. I maintain that it is rather obvious to the reader that anyone who has resolved to undertake a harrowing journey to another state, violating the societal norms of the deep South, and stealing what is legally Mrs. Watson's property does not just have his own interests in mind. A strictly selfish act could have taken the shape of the hypothetical scenario outlined above where Huck acts in agreement with duty, rather than from duty. This itself would only be a selfish act nested within the broader scope of a non-selfish act; after all, Huck only runs into the two men on the skiff because he has already undertaken his mission to rescue Jim.

In addition to this textual evidence, it is clear that strict psychological egoists have their own problems to contend with. Mainly, their agenda does not make sense of moral action or moral praiseworthiness full stop. If psychological egoism is true, a startling implication arises; there is no morally relevant difference between the prudent and the truly good grocer, or Madame Bovary's self-interested charity and the good works of Mahatma Gandhi. This is because altruism, or acting for the good of another, simply becomes another means of self-gratification. The only measure of goodness, if we apply the psychological egoist's agenda to its ethical counterpart, ethical egoism, is the extent to which these acts fulfill the agent's self-interest. Accordingly, the very idea of Huck' s action in rescuing Jim being more morally praiseworthy than turning him in is dead in the water. This is an intuitively unattractive conclusion that makes little sense of morality, and I will dispense with it here.

Conclusion

By demonstrating the interconnectedness of different Kantian concepts and presenting relevant textual examples. I have aimed to demonstrate that by freeing Jim, Huckleberry Finn is acting in accordance with the Kantian MDT and thus it cannot function as a sturdy counterexample to the applicability of this thesis.

References

Arpaly, Nomy. (2006) “Moral Worth,” Unprincipled Virtue: An Inquiry Into Moral Agency, Oxford University Press, pp. 67-116.

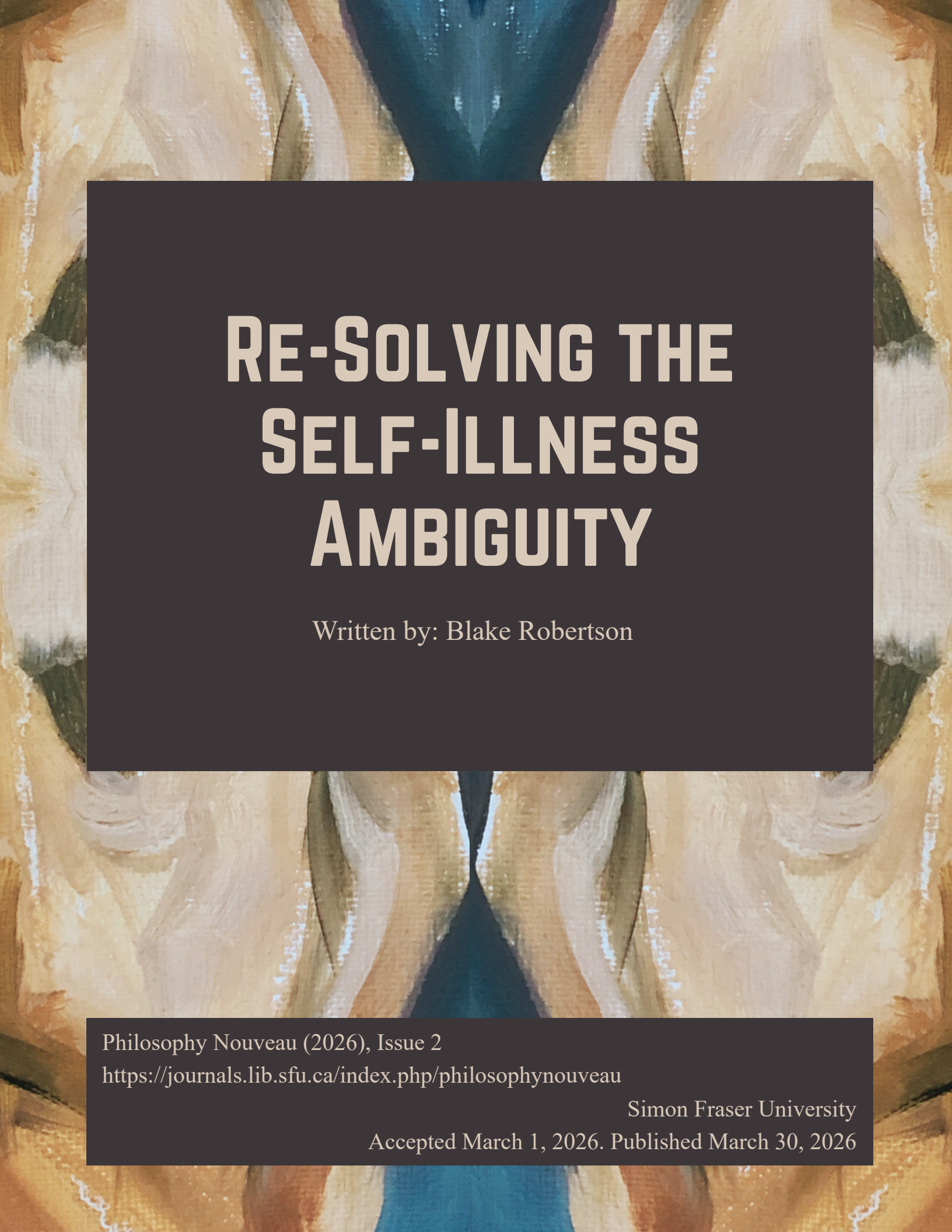
Bennett, Jonathan. (1974) “The Conscience of Huckleberry Finn,” Philosophy, vol. 49, no. 188. Cambridge University Press, pp. 123-134.
Kant, Immanuel. (1785) “Groundwork of the metaphysics of morals,” The Annotated Kant, ed.

Steven Cahn, Rowman & Littlefield Publishers, 2020.

Markovits, Julia. (2010) “Acting for the Right Reasons,” Philosophical Review. Cornell University Publishing, vol. 119, no. 2, pp. 201-242.

Twain, Mark. (1885) “The Adventures of Huckleberry Finn,” Project Gutenberg Ebook Version.

Walschots, Michael. (2022) “Kant and the Duty to Act from Duty” *Penultimate version. Final version published in: History of Philosophy Quarterly, vol. 39 no. 1, pp. 59–75.



Re-Solving the Self-Illness Ambiguity

Written by: Blake Robertson

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau
Simon Fraser University
Accepted March 1, 2026. Published March 30, 2026

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau

Abstract

The self-illness ambiguity, the difficulty of distinguishing between the self and the effects of mental illness, remains a core conceptual and clinical challenge in psychiatry. In "Solving the Self-Illness Ambiguity," Sofia Jeppsson rejects the Realist view that a pre-existing boundary between self and illness can be discovered, proposing instead a Constructivist model in which patients define this boundary in ways that best promote recovery. However, Şerife Tekin critiques Jeppsson's solution for overestimating patients' narrative capacities, ignoring conflicting narratives, and offering little procedural clarity.

In response, this paper proposes a scalar model of agency drawn from Jennifer Hornsby's work on action explanation. Instead of treating agency as all-or-nothing, Hornsby treats it as existing in degrees, shaped by how alienated or involved a person feels in their own actions. I refine this model through three interrelated dimensions: normative engagement, reason-responsiveness, and identification or avowal. Rather than requiring a fixed boundary between self and illness, this model invites patients to assess their degree of alienation from particular actions along these axes.

I support this approach through philosophical and psychiatric literature, drawing on Gloria Ayob, Richard Moran, and Judit Szalai. I also address Tekin's concern about narrative complexity by incorporating Gerrit Glas's model of layered self-referentiality, which allows for partial and evolving self-understanding. Finally, I argue that focusing on bounded, momentary actions, rather than coherence across a life narrative, offers a more realistic and clinically useful framework for navigating self-illness ambiguity.

Simon Fraser University
Accepted: March 1, 2026. Published: March 30, 2026

Introduction

In a clinical reflection, psychiatrist Norman Greenberg dwells on a recent session: "As soon as they left my mouth, the words didn't seem quite right. 'It's Lilly's illness speaking, not her,' I tried to explain to Lilly's partner, who was struggling to understand her newly intensified suicidal behaviour." Greenberg tries, and fails in his view, to beneficially demarcate between a patient's illness and their self. He later ponders "a single unanswerable question": "'If this was Lilly's illness, then which parts were Lilly?" (Greenberg, 2023).

The question of where the self ends and illness begins haunts the field of psychiatry. When a person's illness interferes in their expression of agency, how should we make sense of this alienation? This is the problem of self-illness ambiguity, and its resolution has far-reaching implications for how we assign responsibility, administer care, and understand personal identity in the face of mental disorder (Tekin, 2022; Glas, 2023). With a viable solution still missing (Dings & De Bruin, 2023), I draw on Jennifer Hornsby's theory of action explanation, a theory that critiques a rigid, naturalistic structure that parallels the one Greenberg struggles against. I argue that Hornsby's framework shifts the focus away from fixed self-illness boundaries and toward varying degrees of alienation. From this emerges a scalar model of agency, assessed through three interrelated dimensions: normative engagement, reason-responsiveness, and identification with one's motivations and actions. By grounding this model in contemporary philosophical and psychiatric literature, I aim to provide a more precise, usable, and compassionate approach to understanding agency in mental illness.

Section 1: Jeppsson's Constructivism and Tekin's Criticisms

Jeppsson's Critique of Realism and Advocation for Constructivism

A potential answer to self-illness ambiguity, dubbed the Realist solution, asserts

that there exists a pre-existing, discoverable line between self and illness. The idea is that a psychiatric patient would "discover" this line through reflection and dialogue with their clinician, after which it would remain fixed throughout their life. Actions or behaviours would thus be categorized objectively as either part of the self or the illness.

Sofia Jeppsson, in her paper "Solving the Self-Illness Ambiguity", argues that this solution is both practically untenable and metaphysically unlikely. Jeppsson first argues that it would be futile in practice to attempt to discover this pre-existing line, as patients would inevitably fall into an endless loop of trying to sort their beliefs and behaviours into one of the two categories. Jeppsson uses an example to illustrate this process: a patient with bipolar disorder tries to determine whether their desire to write a novel is genuine or a result of their mania. Each new thought is doubted in the same way, with the patient repeatedly questioning "maybe my previous thought was just a mental illness symptom" (Jeppsson, 2022).

Putting aside the practical feasibility, Jeppsson argues that there is not even a good reason to believe in a naturalistic demarcation. She does so by considering three different theories of self: Sadler's five-aspect self, deep self theories, and the counterfactual self. For the former two, she determines that they offer no clear way of separating out mental illness from features of the self, such as one's agency or core values. For the latter, the counterfactual self, she notes that illness-related counterfactual statements, such as "If I wasn't bipolar, I wouldn't have hastily cut off my friends" are no less valid than everyday counterfactual statements, such as "If I wasn't a foodie, I wouldn't have spent so much on tomatoes." And we obviously do not attempt to demarcate between our selves and our love of food.

In the place of the Realist solution, Jeppsson proposes the Constructivist solution, which involves agents constructing their own boundary in a way that best promotes well-being and function. Jeppsson argues for this solution by appealing

to the central role that narrative authorship plays in patients' interpretations of their illness and self. While Jeppsson's negative argument involves an appeal to the metaphysical impossibility of a fixed boundary, her positive argument for Constructivism is largely instrumental. The standard for success in constructing the boundary is pragmatic: the "best" boundary is the one that maximizes recovery.

Tekin's Critiques of Jeppsson's Constructivism

Şerife Tekin, in "My Illness, My Self, and I", affirms Jeppsson's positive argument while critiquing her Constructivist solution for its idealistic view of narrative authorship, its ignorance of the multiplicity of narratives, and a lack of a procedural clarity. She argues that Jeppsson overly idealizes the capacity for people to construct a coherent self-narrative, let alone people who are faced with one or even multiple mental illnesses. If people with severe disorders like schizophrenia could calmly and rationally tell themselves "that delusion isn't real, it's just my illness", then we wouldn't need psychiatric treatment! The challenge of narrative authorship is further exacerbated by the presence of multiple, oftentimes competing narratives. A person must contend with their own autobiographical narrative along with a variety of social narratives, each of which are likely to feature contradicting beliefs and values. Finally, Tekin notes the lack of procedural clarity in Jeppsson's paper, where she provides no clear guidance for patients to engage in narrative authorship. Jeppsson's solution is no more effective at answering the core question of the self-illness ambiguity: How can patients and clinicians effectively demarcate between their self and their illness? What is needed, Tekin suggests, is a model that preserves the spirit of narrative construction while offering a more grounded account of how agency can persist even when self-narratives fragment or fail.

Section 2: Hornsby's Account of Agency

Hornsby vs The Standard Story

To work towards this model, I begin with a step into the philosophical study of action-explanation and agency. Jeppsson's attack on the Realist solution mirrors a philosophical debate that waged over two decades prior, between philosopher Jennifer Hornsby and the dominant model of action explanation: the standard story. The standard story, espoused by philosophers such as Donald Davidson and J. David Velleman, explains intentional action as the result of a causal chain: a belief-desire pair causes a bodily movement. For instance, my desire to reduce muscle soreness, and my belief that stretching will reduce soreness, causes me to stretch. Hornsby argues that this picture distorts what it means to act by treating agents as sites of causal convergence, rather than evaluative participants in their actions. In curtailing the role of agency in action-explanation, the standard story overly generalizes and too strictly bounds explanations of our behaviours, while simultaneously failing to capture our intuitions about action explanation.

In contrast, Hornsby proposes a model in which agency is not all-or-nothing, and not reducible to internal causes. Instead, she treats agency as graded, embodied, and fundamentally evaluative. A person may act with full confidence and control, or hesitantly, half-heartedly, or in tension with themselves, and yet all of these can still count as agentive actions. The crucial question, for Hornsby, is not whether the right internal states were activated, but whether the person can later make sense of what they did from within their own evaluative perspective. An agent may feel alienated from their action, disoriented, ashamed, or disconnected, and yet still count as the one who acted, albeit with a diminished or fractured sense of agency.

Hornsby's Debate Parallels Jeppsson's

Hornsby's philosophical debate structurally mirrors the debate Jeppsson enters

into for the self-illness ambiguity. In both cases, a dominant model assumes that agency can be sharply delineated: the Realist solution holds that a person's beliefs and behaviors can be objectively sorted into "self" or "illness," while the standard story of action insists that intentional actions can be causally traced solely to belief-desire pairings, and that agency thus exists only in a black-and-white fashion. In both cases, the critics, Jeppsson and Hornsby, argue that such clarity is not only implausible, but unhelpful. Jeppsson points out that patients often experience their symptoms and values as deeply entangled, while Hornsby emphasizes that real-world actions are rarely clean expressions of coherent intention. Both propose more flexible alternatives: Jeppsson through narrative construction, and Hornsby through a model of agency that allows for degrees of self-involvement, hesitation, and alienation. Together, these critiques challenge the idea that agency must be fixed, rational, or unified, and open the door to understanding fractured agency on its own terms.

By shifting the focus from causal mechanisms to evaluative intelligibility, Hornsby offers a framework that makes space for the kind of agency Jeppsson and Tekin are circling: one that does not require coherent authorship, total control, or a clear self/illness boundary in order to be meaningful.

Section 3: Using Hornsby's Model to Solve Tekin's Critiques

To restate, Tekin's main critiques of Jeppsson are that she is overly idealistic of patients' ability to construct a coherent self-narrative, ignorant of the multiplicity of narratives that patients must contend with, and lacking in procedural clarity. I will lump together the first two critiques into the following problem: The model that solves the self-illness ambiguity must provide a clear and usable framework for assessing oneself and one's illness. I will solve this problem by introducing a scalar form of agency for assessing one's alienation from one's actions in the face of intrusions from the illness. I will then clarify the meaning of alienation, splitting it into three distinct but overlapping scales: normative engagement, reasons-

responsiveness and identification. In later sections, I will ground this model in philosophical and psychiatric literature, consider an objection, and finally respond to Tekin's second critique.

A Scalar Sense of Agency

Using Hornsby's scalar understanding of agency, we can make an immediate amendment to Jeppsson's constructivist model: patients should not attempt to cleanly demarcate between their illness and their self, but rather should assess the degree to which they feel alienated from their behaviour as a result of their illness. For instance, a patient with anxiety may feel that their illness has less of an effect on their social abilities when they are around close friends or engaged in an activity that they are skilled at. This change, to a scalar model of agency, lowers the bar for participation in narrative self-authorship: patients do not need to make black-and-white assessments of their actions and behaviours, and can better interpret their illness as interwoven into their self, not as something to be cleanly demarcated out.

Fixing Up Hornsby to Fix Up Jeppsson

Unfortunately, this still leaves us with a significant problem: how are agents supposed to assess their degree of alienation? For this, we turn again to Hornsby. As stated, Hornsby thinks of agency as existing in degrees, with the following factors determining where someone lies along that scale:

> *Our conception of an agent-in-the-highest-degree might be a conception of someone who is fully self-reflective and* ***has complete self-control****, who has values and* ***makes valuational judgements*** *upon which she acts, who* ***uses reason and argument effectively****, who is sensitive to her circumstances, who puts her heart into what she does, and who, as we say,* ***identifies with her motivations and with what she***

> ***does***. *To the extent to which a person's doing something on an occasion shows her as deficient compared to an agent-in-the-highest-degree, we could think of her as failing to participate in Velleman's agency par excellence. This is now to think of her as falling short of some ideal or other, and not as lacking some causally potent brain state.* Hornsby, Agency and Alienation

At the root of Hornsby's argument here, and throughout "Agency and Alienation", is the eponymous belief that degrees of agency exist as a scale between "full-blooded agency" and "complete alienation." As stated in the above quote, Hornsby envisions an agent's placement on this scale as being determined by their degree of:

1. Self-reflection and self-control
2. Possession and use of normative beliefs
3. Reason-responsiveness
4. Sensitivity to their circumstances
5. Heart put into what they do
6. Identification with their own motivations and actions

To better determine how agents assess the degree to which their illness has affected their behaviour, I want to narrow down these categories. We can begin by removing number 5, "Heart put into what one does." It feels too vague and storybook-esque, not too mention redundant in the face of the nearly identical category "Identification with their own motivation and actions." Next, we move on to "self-reflection and self-control." Self-reflection can easily be consumed by the aforementioned category "identification with their own motivations and actions", as to be self-reflective is merely to assess one's own present and past beliefs and actions, and I see no reason why identification is not something that can be backwards looking. Self-control seems similarly consumable by "possession and use of normative beliefs" and "reason-responsiveness", as the act of controlling

oneself seems rooted in the appeal to values and reason. Finally, "sensitivity to their circumstances" is subsumed under the remaining three categories, as it is simply an outward looking form of self-reflection. Thus, after narrowing down Hornsby's categories, we are left with the following three for assessing an agent's degree of alienation:

1. Possession and use of normative beliefs
2. Reason-responsiveness
3. Identification with their own motivations and actions

A scalar understanding of agency, alongside the above categories, now gives us a method for assessing the degree of effect that a patient's illness has on their behaviour. Patients no longer have to firmly demarcate between their illness and their self with some vague guiding principle of "improving recovery." In assessing their own self-illness ambiguity, agents should attempt to comprehend their degree of alienation from their actions. Even this assessment tool is still too vague though, and for that we turn to the narrowed down version of Hornsby's determiners of agency, with the three categories outlined above.

Section 4: Grounding, Clarifying, and Contradicting

To properly argue for the value of these three categories in solving the self-illness ambiguity, I must ground them in additional philosophical and psychiatric literature. I will do this by drawing a comparison between my first two dimensions of agency and dimensions that Gloria Ayob discusses in "Agency in the absence of reason-responsiveness." Next, I will draw a similar parallel, but this time between my third dimension and the concepts explicated in Richard Moran's "Authority and Estrangement." Finally, I will consider and resolve a counterexample provided by Judit Szalai in "Agency and Mental States in Obsessive-Compulsive Disorder."

Grounding in Ayob

In “Agency in the absence of reason-responsiveness”, Ayob primarily concerns herself with responding to and resolving an argument made by fellow philosopher of psychiatry, Hanna Pickard. Pickard argues that agency should be understood in terms of choice and control, and that we should divide behaviour into two kinds: mere bodily movements (e.g., automatic reflexes) and action, which is voluntary by definition. This binary distinction is highly reminiscent of the black-and-white thinking with which Velleman approaches the concept of agency in action-explanation. The parallel continues, with Ayob responding to Pickard in a distinctly Hornsbian fashion: She advocates for a graded spectrum of agency, between minimally voluntary action and substantively voluntary action.

These two ends of Ayob’s spectrum are distinguished (note, they are not demarcated, as there is no firm boundary) based on an agent’s “engagement in reason-giving practices” (Ayob, 2016). Specifically, Ayob envisions that these are reasons that possess “normative pull.” For instance, an agent who drives a car drunk, despite latently knowing the risks to themselves and others, fails to engage with the normative reasons that oppose their behaviour. Possessing normative beliefs and responding to reasons are not just psychological states, they are markers of the agent’s participation in normative practices. This allows us to distinguish between a person who simply acts in a goal-directed way, and one who acts for reasons they can endorse. These two factors, reasons-responsiveness and normative engagement, directly mirror the first and second categories that make up my modified version of Hornsby’s account of scalar agency.

This modern, psychiatrically-inclined parallel that Ayob offers is not her only contribution, however. Ayob introduces an important distinction between two different forms of reasons-responsive behaviour: reason-blindness and reason-insensitiveness. A person who is reason-blind fails to appraise reasons properly,

whereas a person who is reason insensitive manages to appraise reasons properly, but fails "to be appropriately motivated by those reasons" (Ayob, 2016). Both forms of lacking reasons-responsiveness are present in illness-related behaviour. For instance, a person with anxiety may fear that everyone will laugh at them for a minor social infraction (reason-blind), whereas a person with ADHD may struggle to do their homework despite being acutely aware of its importance (reason-insensitive). Thus, the distinction is key because it gives us an additional scale to assess an agent's degree of illness-related alienation from their actions.

Grounding in Moran

Moran serves a similar role to Ayob, with his framework reinforcing the scalar model of agency by showing that even when a person acts for reasons (and can reflect on those reasons) they may still feel estranged from the action. In "Authority and Estrangement", Moran distinguishes between two forms of first-personal epistemic engagement with one's mental states: avowal and attribution. To avow a mental state (e.g., a belief, a desire, an intention) is to express it as one's own, from the first-person standpoint. Avowals are acts of self-commitment to the epistemic and/or ethical value of such a mental state. On the other hand, to attribute a mental state is to describe it as a fact about yourself from a more external perspective, recognizing or diagnosing your own state rather than speaking from within it.

Moran argues that true agency requires avowal, not mere attribution. This fact, along with Moran's definitions, strongly parallel my third category "identification with their own motivations and actions", while also further developing our assessment of patients through the lens of that third category. To avow, or identify with your motivations and actions is to see yourself represented in them, to support their truth and moral worth as propositions. It is not merely to recognize them, or to be in a state to have them coaxed out of you. Moran clarifies this distinction with an example: a psychiatric patient is made aware of her belief that

she is angry at her dead parent for "abandoning" her. While the patient recognizes that this belief is not grounded in reality, she can still attribute the belief to herself, and thus she does not exercise complete agency in acting in relation to the belief.

This example draws out a key point that Moran himself does not make: avowal and attribution, and the agentive actions associated with them, can occur in degrees. For instance, the patient who merely attributes the anger spurred by their dead parent's "abandonment" may genuinely avow the belief that their parent should have driven less recklessly, or taken better care of their health (or whatever decisions relate to the cause of death). Avowal and attribution can coexist and even blur within the same emotional experience. Thus, actions taken in relation to an emotional experience can be more or less agentive based on the relative presence of avowal and attribution in its makeup. Moran thus helps to further ground my Hornsbian model in psychiatric-adjacent philosophy, while clarifying the criteria by which we judge patients' identification with their own motivations and actions.

Briefly Clarifying the Model

Before considering an objection, I want to synthesize the ideas just presented by clearly defining the three interrelated dimensions by which I argue we should assess agency in psychiatric contexts:

1. Normative Engagement: The degree to which the agent's action is guided by, or responsive to, reasons with ethical or evaluative weight. I.e., not just goal-seeking, but value-sensitive.
2. Reason-Responsiveness: The capacity to properly appraise and be appropriately motivated by reasons, both internal and external.
3. Identification (Avowal): Whether the agent sees the action or motivation as theirs, in the sense of being able to avow it from a first-person standpoint. This includes taking responsibility for the intention, motivation, or outcome of the action, and being able to assess one's motivations in a present and forward-

looking manner.

Contradictions in Szalai

Now that we have fully synthesized our Hornsbian model of agency, we move on to consider an objection. In “Agency and Mental States in Obsessive-Compulsive Disorder”, Judit Szalai generates a counterexample that is eerily fitting to our multidimensional model of agency. Szalai argues that OCD compulsions are agentive actions because they are voluntary and goal directed, despite them often lacking reasons-responsiveness, normative endorsement, and avowal. In other words, we have a case of a set of actions that score low on all three categories of our Hornsbian model, and yet still seemingly count as agentive.

A closer look, however, reveals that OCD compulsions, like many psychiatric behaviors, do not represent an absence of agency but rather a distorted or conflicted expression of it. Consider the axis of normative engagement: while the compulsion itself may not be endorsed, the underlying goals, such as avoiding harm or reducing anxiety, are often normatively compelling to the agent. The person may not want to wash their hands for the twentieth time, but they do value safety, responsibility, and the reduction of anxiety. Similarly, the compulsion is often grounded in subjectively intelligible reasoning, even if that reasoning does not hold up to objective scrutiny. The patient acts on what feels like a reason, even if they recognize it as irrational. Finally, while the agent may disavow the behavior at a reflective level, they often acknowledge their motivation: “I had to do it to calm down,” or “It was the only way I could sleep.” This is not full avowal, but it is not pure attribution either: it reflects a partial identification with the action’s motive, if not its form.

Thus, Szalai’s example does not falsify the scalar model, it confirms its necessity. OCD compulsions are not simply “non-agential”, they are complex, layered actions that sit low on the scalar spectrum but remain intelligible as products of a

fractured agency. These are precisely the kinds of cases the model is built to address: actions that are neither fully voluntary nor fully passive, but somewhere in between. What Szalai helps illuminate is not a weakness in the scalar model, but its strength, its ability to accommodate the murky, ambivalent, and often painful realities of psychiatric life without resorting to binary categories of agency versus illness. The competing realities present in OCD compulsions brings out a key point that must now be handled: the challenge of the multiplicity of narratives.

Section 5: The Multiplicity of Narratives

So at last, we deal with the second of Tekin's primary critiques of Jeppsson: her ignorance of the multiplicity of narratives. Tekin is concerned that Jeppsson does not properly recognize the challenge associated with synthesizing autobiographical and social narratives to form a coherent self-story, especially because both narratives (particularly the latter) are likely to harbour contradicting propositions about the agent's life. There is no way to cleanly diffuse this criticism, as the very nature of self-assessment involves a variety of perspectives (past and present) intervening on your beliefs. Thus, we must look for a solution that allows us to lower the demand for self-understanding while making room for narrative conflict.

Glas's Levels of Self-Referentiality

Such a model is proposed by Gerrit Glas in "Dimensions of self-illness ambiguity", wherein he offers a way to diffuse the pressure of competing narratives by demonstrating that narrative coherence is not a prerequisite for meaningful self-understanding. His concept of layered self-referentiality distinguishes between three ways people relate to themselves in the context of illness: primary self-referentiality, which involves the raw, immediate feeling of being unwell or disconnected from oneself; secondary self-referentiality, which concerns how people emotionally or practically respond to their illness in daily life; and tertiary

self-referentiality, which involves explicit reflection and narrative interpretation.

Crucially, Glas does not assume that these levels must always be active together or in harmony. A person might experience deep alienation at the primary level, manage their illness adaptively at the secondary level, and still lack a coherent narrative at the tertiary level. Or they may have moments of reflective insight that are not yet integrated into their everyday habits or affective experience. What this model provides is a philosophically grounded way of understanding partial, fragmented, or developing forms of agency. It allows patients to make sense of themselves incrementally, rather than all at once, and to occupy positions of uncertainty without forfeiting agency altogether. This reframes the multiplicity of narratives not as a breakdown of authorship, but as a natural and even expected condition of living with mental illness. Glas doesn't eliminate the narrative tension Tekin points to, but he shows that it does not preclude meaningful participation in one's self-understanding. In doing so, he strengthens the scalar model of agency proposed in this paper by adding a temporal and structural account of how people relate to themselves over time, across different layers of experience.

These layers of experience could also be separated in treatment to better address contradicting beliefs within one's autobiographical narrative. For instance, an agent could harbour extreme grief that is expressed through their body language, energy levels, facial expressions, or appetite (primary self-referentiality). The agent could respond to this grief by throwing themselves into work and hobbies (secondary self-referentiality), thus appearing mentally well, and maybe even reflectively convincing themselves that they do not need support (tertiary self-referentiality). A clinician, or even the patient on their own, can more effectively respond to the contradictions present in their autobiographical narrative by separating each level of self-referentiality.

A Lesson from Action-Explanation

While Glas's model helps to manage some of the struggles present in the multiplicity of narratives, its ability to resolve tension is largely relegated to the autobiographical sphere. To handle the presence of contradiction in social narratives is far more challenging, for the clinician has only secondary access to these narratives, and may struggle to cleanly demarcate between the agent's own beliefs, the social narratives that colour these beliefs, and the social narratives that are floating around, bouncing against the agent's own beliefs. To help to resolve this challenge, I point to a practice consistent through the disciple of action-explanation: to focus on individual, momentary experiences in assessing agency, alienation, beliefs, desires, and any other concept of action-explanation.

Hornsby, Davidson, and Velleman do not speak of broad, overarching self-narratives when explicating their respective theories, instead they describe momentary actions: pressing a brake pedal, lifting one's arm, flicking a light switch. This decision, to bound narratives, is made because it makes it plainly easier to make narrative and metaphysical assessments. Clinicians and patients can take a page out of this book, as by assessing individual actions or beliefs as they come up, there is simply fewer social narratives that either party is forced to contend with.

This bounding of narratives also better aligns with the constructivist model as a whole. As situations continue to develop and patient's emotions change over time, patients can alter their degree of alienation in a way that maximizes recovery. If it is better for a patient to tell themselves that their workaholic tendencies are an agentive choice (and not one made out of grief) during the early stages of loss, the patient has the choice to alter their agentive relationship to the action if they later feel it is beneficial to alienate themselves from it. The bounding of narratives, from overarching stories of behaviours to individual actions, allows patients and clinicians to assess events with a reduced presence of social narratives.

Additionally, the evolving nature of momentary experiences better aligns with the temporally continuous nature of the constructivist solution.

Conclusion

The self–illness ambiguity demands a theory of agency that can live up to the complexity of psychiatric life, one that does not collapse under the weight of its own binary assumptions. In this paper, I have argued that Jennifer Hornsby's scalar account of agency provides the tools needed to refine Jeppsson's constructivist model and respond to Tekin's core critiques. By analyzing agency through the dimensions of normative engagement, reason-responsiveness, and avowal, we can move beyond idealized notions of authorship and begin to make sense of fractured, partial, or conflicted self-understanding.

What emerges is not a clean map of where illness ends and self begins, but a flexible framework for tracing where and how agency shows up: in actions that are endorsed or regretted, intelligible or compulsive, owned or disowned. With support from Ayob, Moran, Szalai, and Glas, the scalar model of agency makes room for ambiguity, without abandoning accountability. It reframes psychiatric identity not as a question of drawing a boundary, but of navigating one's position along overlapping continua of selfhood, motivation, and meaning. In doing so, it provides a more realistic and ethically sensitive foundation for understanding agency in the context of mental illness.

Acknowledgements

Thank you to Dr. Holly Anderson for talking me out of a few bad ideas and into a couple of good ones. Thank you to my research cluster: Nida Anwar, Gerard Corr, Shae Sackman, and Eduardo Stehling, for motivating and sharpening my work. Thank you to Nava Karimi for destroying my initial, much worse argument. Thank you to Isaac Malmgren for kindly redirecting me when my argument got off track. And thank you to Jaskaran Rai for poking a very important hole in my argument.

References

Jeppsson, S. M. (2022). Solving the self-illness ambiguity: the case for construction over discovery. Philosophical Explorations, 25(3), 294-313.

Tekin, Ş. (2022). My Illness, My Self, and I: when self-narratives and illness-narratives clash. Philosophical Explorations, 25(3), 314-318.

Hornsby, J. (2008). Agency and alienation.

Hornsby, J. (2004). Agency and actions. Royal Institute of Philosophy Supplements, 55, 1-23.

Davidson, D. (1963). Actions, Reasons, and Causes. The Journal of Philosophy, 60, 85-700

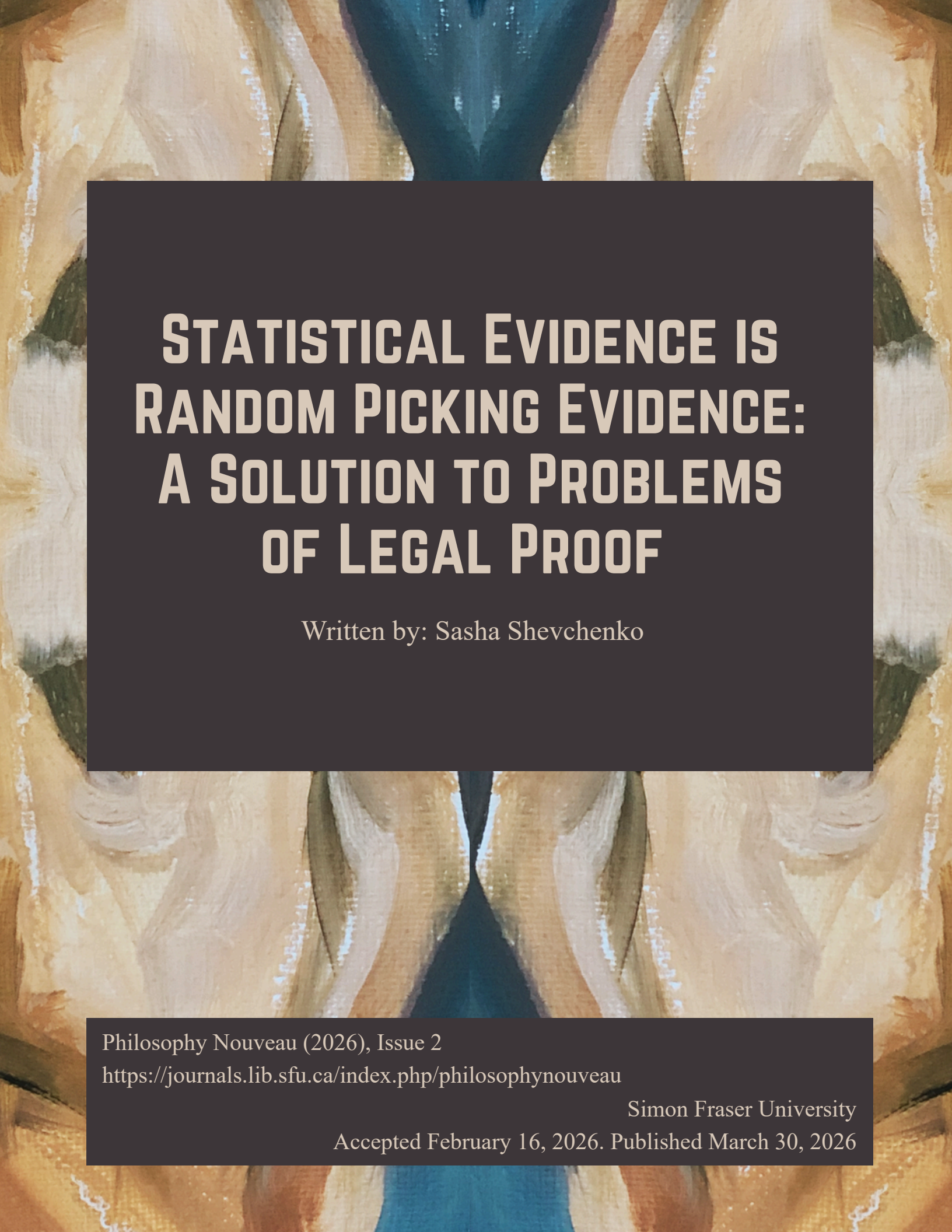
Dings, R., & De Bruin, L. C. (2023). self-illness ambiguity and narrative identity. Philosophical Explorations, 26(2), 147-154.

Ayob, G. (2016). Agency in the absence of reason-responsiveness: The case of dispositional impulsivity in personality disorders. Philosophy, Psychiatry, & Psychology, 23(1), 61-73.

Szalai, J. (2016). Agency and mental states in obsessive-compulsive disorder. Philosophy, Psychiatry, & Psychology, 23(1), 47-59.

Glas, G. (2023). Dimensions of self-illness ambiguity–a clinical and conceptual approach. Philosophical Explorations, 26(2), 165-178.

Greenberg, N. R. (2024). When the illness speaks. BJPsych Advances, 30(3), 164-165.



Statistical Evidence is Random Picking Evidence: A Solution to Problems of Legal Proof

Written by: Sasha Shevchenko

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau

Simon Fraser University

Accepted February 16, 2026. Published March 30, 2026

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau

Abstract

A puzzle arises from the intuitive idea that it is a mistake to find someone liable merely because purely statistical evidence indicates that the probability the offence was committed meets a standard of proof. The proof paradox is the intuition that basing legal verdicts on individual evidence is appropriate while basing them on bare statistical evidence is not. One explanation is that "individual evidence is about the individual in a way that statistical evidence is not." (Enoch et al. 5) I aim to vindicate this explanation by making precise the way in which bare statistical evidence fails to be about the individual. I propose that the notion of random picking is the criterion that distinguishes bare statistical evidence from individual evidence.

Simon Fraser University
Accepted: February 16, 2026. Published: March 30, 2026

Statistical Evidence is Random Picking Evidence
Sasha Shevchenko

Introduction

A particular type of puzzle arises from the intuitive thought that it is a mistake to find someone liable merely because purely statistical evidence indicates that the probability the offence was committed meets the standard of proof. The proof paradox is the problem of explaining why we ordinarily think that basing legal verdicts on individual evidence is appropriate while basing them on bare statistical evidence is not, even when both are of equivalent quality. One prima facie explanation is that "individual evidence is about the individual in a way that statistical evidence is not." (Enoch et al 5) I aim to vindicate this intuition by making precise the way in which bare statistical evidence fails to be about the individual, and how this makes its use objectionable. To do so, I will propose that the notion of random picking is the criterion that distinguishes bare statistical evidence from individual evidence. The supposed paradox will be shown to arise from the ways in which the involvement of random picking is (to a greater or lesser degree) obscured by the structure of the examples.

The distinction that I suggest demands a positive characterization of what bare statistical evidence really concerns, one that would shed light on how it fails to properly connect to particular individuals. To provide such a description, I draw on the principle of Random Picking: (RP) The fact that a high proportion of Fs are Gs cannot, in and of itself, normically support the proposition that the result of randomly picking an F is a G. (Blome-Tillman 569) This principle captures both what makes bare statistical evidence bare, and why this is so problematic. Statistical evidence of the kind deployed in (RP) is about Fs and Gs as categories; it makes no reference to any specific individuals or their properties. The problem arises from the fact that the arbitrariness of the selection method precludes the need to explain the features of the result. A crooked die may roll 5s 90% of the time, but the fact that it rolled a 4 this once cannot be explained by statistical facts. All the statistics can say is that in 10% of cases, the roll will not result in a 5, and this was one of those times. Only the particular features of the roll could tell us anything

about why this 10% was instantiated by this roll and not another. This is exactly the information that bare statistical evidence does not supply. Thus, the essential feature that makes reasoning from bare statistical evidence problematic is the fact that it cannot be used to support conclusions about arbitrary particulars. The resolution to the problem of statistical evidence I propose embodies this fact: if (RP) is present, the probabilistic evidence used in the antecedent is bare statistical and cannot sustain verdicts.

To demonstrate how random picking resolves proof paradoxes, I will consider the paradigmatic Blue Bus/Red Bus case. In this scenario, the victim is struck by a bus that they cannot identify. We are asked to consider two potential evidential bases for a finding of liability (Enoch et al. 197). The first basis is eyewitness testimony; suppose the eyewitness testifies that they saw a Blue Bus hit the victim, and that they are reliable enough that there is a 0.7 probability that they have correctly identified Blue Bus as liable. The second basis is market share evidence. Here, we know only that 70% of the buses in the area near the incident are operated by Blue Bus. The rest are run by their competitor, Red Bus. This entails that there is a 0.7 probability that a Blue Bus really did strike the victim, and that the company is liable. We must then account for why a finding of liability on the basis of this bare statistical evidence seems mistaken, whereas the eyewitness testimony appears perfectly sufficient (Enoch et al. 198-199). Applying the (RP) standard cleanly resolves this issue. The bus that struck the victim is nothing more than a single random sample of the buses that operate in the area. The market share evidence only shows that, if we were to select a random bus, we would find it is a Blue Bus 70% of the time. But the selection process has already occurred; the bus has already struck the victim. Thus, the fact that a high proportion of buses are Blue Buses does not normically support the proposition that the bus that struck the victim is a Blue Bus. So the market share evidence runs afoul of the (RP) criterion. And since the eyewitness testimony is directly connected to the features of the result of random picking – the bus that actually struck the victim – (RP) is not implicated there. As such, (RP) provides a clear, bright-line distinction between the

bare statistical and individual evidence cases, just as our intuitions demand. Since the Blue Bus/Red Bus case is the prototypical example of a proof paradox, (RP)'s success makes it plausible that it will be able to capture all problems of bare statistical evidence.

However, new cases could be constructed which do not involve a random selection method. There are two basic ways of making a selection procedure non-random in the legal context: either the choice is based on individual evidence, or on bare statistical evidence. If the evidence used against the accused is individual – such as eyewitness testimony – then the (RP) standard will indeed not be violated, as discussed above. But that is precisely the desired result; the use of (RP) is meant to classify the deployment of individual evidence as acceptable. So whenever we find that a putative counterexample to (RP) really involves the use of individual evidence for non-random picking, we can immediately conclude that this case is unproblematic. Only cases that contain both bare statistical evidence and non-random picking pose a threat to (RP). This is because it would appear that (RP) would deem the verdicts in such cases acceptable, contrary to our intuitions about the individual-statistical distinction.

While originally meant as an argument against the normic support standard, Blome-Tillman develops a counterexample of precisely this kind in the form of the Political Gatecrasher. He gives the scenario as follows:

The Political Gatecrasher: *The organizers of the local bullfighting decide to sue Luis for gatecrashing their Sunday afternoon event. Their evidence is as follows: Luis attended the Sunday afternoon event—he was seen and photographed on the main ranks during the event. No tickets were issued, so Luis cannot be expected to prove that he bought a ticket with a ticket stub. However, while 1,000 people were counted in the seats, only 300 paid for admission. Flyers by anonymous anti-bullfighting activists were found in the arena claiming responsibility for the gatecrashing. Luis is a 22-year-old political science student, and belongs, as such,*

to a group of people who are extremely unlikely to attend a bullfighting event under ordinary circumstances. (Blome-Tillman 570)

In this case, the fact that young political science students are highly unlikely to attend the event without gatecrashing serves as the bare statistical evidence used to pick out Luis as the defendant. Blome-Tillman is explicit in saying that the reason this picking is non-random is that it is predicated on this bare statistical evidence about Luis' demographic membership. (Blome-Tillman 570) The way in which Blome-Tillman contrasts this case with a nearby, non-problematic version serves to illustrate why he understands the Political Gatecrasher this way – and so, why it is potentially threatening for (RP). He claims that were Luis selected and charged randomly, and it just so happened that he is a young political science student, this would be a straightforward violation of the (RP) standard (Blome-Tillman 572). I thus interpret Blome-Tillman as asserting that what makes the picking in the Political Gatecrasher non-random is that Luis is chosen from among the rodeo attendees on the grounds of a particular feature about him – namely, his demographic membership. This is itself a direct consequence of another innovation. The bare statistical evidence in this case is connected to Luis in a way that typical (RP)-violating evidence is not. That he is part of a group that is highly unlikely to have attended the event without gatecrashing is a statistical fact about Luis. In this respect, it is disanalogous with market share-style evidence, which is a statistical fact about a category of individuals writ large – all buses operating in a given area, say. These are what I take to be the strongest considerations that would motivate an objector to join Blome-Tillman in pressing the point that "intuitions about the fact that the results of random pickings are not in need of explanations are thus irrelevant with respect to the Political Gatecrasher." (Blome-Tillman 572)

In the face of these issues, it is clear that defending the (RP) standard requires that we find an instance of random picking in this case. It is equally clear that this cannot be done by examining Luis' selection as the accused from among the

attendees; I believe that Blome-Tillman's reasoning on this particular point is well-founded. Therefore, the only strategy open to the defender of (RP) is to argue that the problem of random picking has not been solved, but merely pushed back – that is, to show that there is another, unrecognized random picking located elsewhere in the example. This is the strategy I will pursue.

Luis is selected randomly in the sense that he is a random sample of young political science students. Nothing in this case distinguishes him from any other member of this demographic. While it may be true that it is highly unlikely that a person like him would go to a bullfighting event, Luis may very well be one of the few who have no qualms with such things. Supposing that he is, there is nothing to explain about his innocent presence at the event, and a great deal to explain about why he would gatecrash. This line of reasoning can be readily summarized as follows:

(RPPOL) *The fact that a high proportion of young political science students are unlikely to attend bullfighting events (without gatecrashing) cannot, in and of itself, normically support the proposition that the result of randomly picking a young political science student is also unlikely to attend bullfighting events (without gatecrashing).*

This is a clear, unmistakable instance of the kind of issue singled out by (RP). In employing bare statistical facts about persons, we improperly reason from the general to the particular. The fact that an individual is a member of a group that is likely to commit an offence does not mean that this individual is likely to commit an offence. This is what I believe to be the distinctive irrationality behind profiling. It is precisely this irrationality that is captured in RPPOL, and what I take to ground our intuition that the Political Gatecrasher is an example of the unacceptable use of bare statistical evidence. Thus, it would appear that the non-random selection of Luis from among the audience served to obscure the actual problem feature in this case – the fact that he is a random, arbitrary representative of his demographic group. If this is right, then the Political Gatecrasher case can be fully accounted for

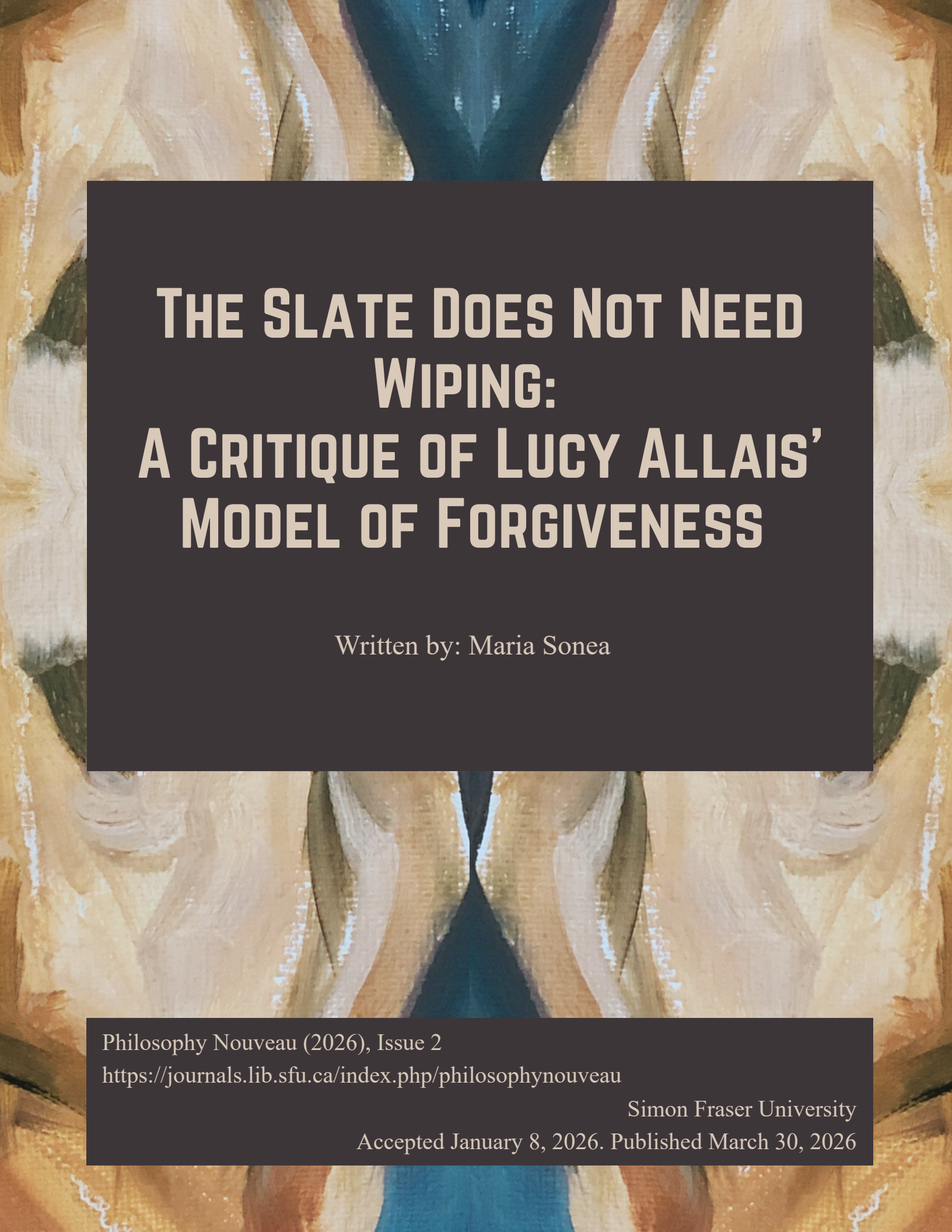
in terms of the (RP) criterion for bare statistical evidence. Moreover, the success of this strategy provides a blueprint for addressing any future challenges developed along similar lines. The procedure is simple: trace the structure of the example carefully, and see if (RP) appears in a disguised, unobvious form.

The distinction between individual and bare statistical evidence requires a systematic criterion that can justify the intuitive refusal to rely on bare statistical evidence in court. (RP) proves to be up to the challenge. Where our intuitions of inappropriateness are at their strongest – in market-share style cases – (RP) provides a simple way to distinguish and condemn the use of bare statistical evidence. Furthermore, it is immune even to the more challenging profiling-style examples such as the Political Gatecrasher. These facts lend credence to the claim that this approach will generalize to account for all apparent proof paradoxes. Thus I consider all problems of bare statistical evidence to reduce to instances of random picking in violation of (RP).

References

Enoch, David, Levi Spectre, and Talia Fisher. (2012) "Statistical evidence, sensitivity, and the legal value of knowledge." Philosophy & Public Affairs, vol. 40, pp. 197-224.

Blome-Tillmann, Michael. (2020) "Statistical evidence, normalcy, and the gatecrasher paradox." Mind, vol. 129, pp. 563-578.



The Slate Does Not Need Wiping: A Critique of Lucy Allais' Model of Forgiveness

Written by: Maria Sonea

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau
Simon Fraser University
Accepted January 8, 2026. Published March 30, 2026

Philosophy Nouveau (2026), Issue 2
https://journals.lib.sfu.ca/index.php/philosophynouveau

Abstract

In “Wiping the Slate Clean: The Heart of Forgiveness,” Lucy Allais attempts to articulate a coherent secular notion of forgiveness. She suggests that forgiveness involves ceasing to hold personal retributive reactive attitudes toward the wrongdoer. This paper critiques Allais’ account, arguing that her definition fails to capture the flexibility inherent in forgiveness. I challenge the tendency to treat affective neutrality as the aspirational goal of forgiveness, rejecting the notion that it must involve ceasing personal retributive reactive attitudes. Harm inevitably alters relationships, and healing occurs through ongoing positive actions from the perpetrator. Forgiveness remains elective, decided solely by the victim, and is never morally obligatory.

Simon Fraser University
Accepted: January 8, 2026. Published: March 30, 2026

The Slate Does Not Need Wiping
Maria Sonea

Introduction

In a 2008 paper entitled “Wiping the Slate Clean: The Heart of Forgiveness,” Lucy Allais attempts to articulate a coherent secular notion of forgiveness. She aims to capture the metaphor for “wiping the slate clean” without altering judgements about the wrongness of an offense or the perpetrator’s culpability. Allais argues that forgiveness centrally involves ceasing to hold personal retributive reactive attitudes toward the wrongdoer, drawing on Peter Strawson’s framework of reactive attitudes to explain how victims can separate the person from the act affectively, while maintaining cognitive judgements of wrongdoing.This paper critiques Allais’ account, arguing that her definition is overly rigid and fails to capture the flexibility inherent in forgiveness. I challenge the tendency to treat affective neutrality as an aspirational endpoint of forgiveness. While I agree that reactive attitudes are interpersonal responses rooted in expectations of goodwill, I do not endorse that forgiveness must involve ceasing personal retributive reactive attitudes. Harm inevitably alters relationships, and healing occurs not through forgiveness alone but through ongoing positive actions from the perpetrator. Forgiveness remains elective, decided solely by the victim, and is never morally obligatory.

I begin by argument by expositing Allais’ paper, providing a detailed explanation of her argument for forgiveness requiring the cease of personal retributive reactive attitudes. Then, I discuss my argument about the flexibility of forgiveness and explain why victims of wrongdoing are never morally required to forgive their wrongdoer. I further test the efficacy of my argument through case studies, examining variations of wrongdoing and relationships. I will include Allais’ South African Truth and Reconciliation Commission (TRC) examples as well as a scenario from the 2014 Argentinian film Wild Tales. I consider a possible objection to my argument before concluding that the elective nature of forgiveness ought not to demand emotional erasure on the part of the victim.

Lucy Allais on Forgiveness

In “Wiping the Slate Clean: The Heart of Forgiveness,” Lucy Allais queries the concept of forgiveness, specifically, how it is possible to “wipe the slate clean” while retaining judgements of wrongness and culpability. (33) She identifies conceptual difficulties, such as ceasing to blame without erasing responsibility, and justificatory issues, like avoiding condonation without atonement (34). Citing Jacques Derrida, Aurel Kolnai, Jeffrie Murphy and Jean Hampton, she distinguishes forgiveness from excusing, justifying, accepting, or ignoring the offender. (34 - 35)

Allais is primarily concerned with the core notion of forgiveness regarding unjustified and unexcused wrongs that warrant retributive censure. (34) She uses the example of a person lying to their jealous partner about visiting with an ex to illustrate this point. (36) Forgiveness, here, is personal, in the sense that the victim is the one who grants it to their wrongdoer, and is incompatible with continued hostility towards the wrongdoer. Allais also describes forgiveness as a gift; it can be withheld or given regardless of whether the wrongdoer has repented. (37) Allais critiques “weak” accounts, particularly theological, where forgiveness overcomes inappropriate resentment once atonement is accomplished, favouring “stronger” views overcoming justified resentment as discussed by David Novitz and Piers Benn. Novitz suggests that forgiveness is expected in situations where our emotions about a wrongdoing are appropriate to the wrong that we believe has been done. (39) Benn reiterates the legitimacy of these emotions that forgiveness depends on. (39)

Allais introduces two real life cases of forgiveness from the South African Truth and Reconciliation Commission (TRC). The first documents Babalwa Mhlauli’s desire to forgive her father’s murderers, asking to know who did it so that she could properly forgive him. The second documents Eugene de Kock’s personal apology to the widows of two of his murder victims and being forgiven by both. Building on Murphy’s account of overcoming retributive emotions, Allais assumes emotions

have intentional content. (41-43) She argues that overcoming emotions alone is insufficient if the wrong still negatively affects the view of the wrongdoer. Forgiveness requires an affective change separating the wrongdoer from the wrongful act.

Overcoming Personal Retributive Reactive Attitudes

Allais draws on Strawson's account of reactive attitudes to more closely define what what constitute forgiveness. She outlines five key features of Strawson's account. First, reactive attitudes are feelings, not beliefs, that enable affective changes without cognitive shifts. (52) Second, they are affective attitudes, which can be understood as distinct from singular emotions like anger. Affective attitudes are ways of feeling towards a given person. (52) Third, reactive attitudes have fitting conditions, which involve participating in relationships in such a way that one responds to the way others manifest their will through their actions, embodying "recognition respect," which Stephen Darwall describes as something "we owe to all rational/moral agents, in virtue of their being agents, regardless of how well they act." (53) Fourth, our reactive attitudes reflect a demand for people to demonstrate goodwill to us and respond to the way people manifest good or ill will towards us. (54) It is here that Allais suggests reactive attitudes involve "esteem respect." (53) in addition to recognition respect. Esteem respect can be understood as admiring a person's worthiness based on their actions. (52) If a person harms or wrongs another person through their action, that action will alter the victim's feelings towards the wrongdoer either by reducing trust or inducing contempt, even if anger fades. Fifth, Allais focuses on personal retributive reactive attitudes, which entail that forgiveness is given to the wrongdoer by the victim of the wrongdoing. (55) Retributivism in this context refers to the idea that wrongdoing should be proportionately censured.(55) Forgiveness, under Allais' view, is a changing of these attitudes without changing judgement of the wrongdoer's responsibility. She states that forgiveness "results in a changed view of the wrongdoer as a person in which you cease to have towards her the personal

retributive reactive attitudes that her wrongdoing supports." (57) By specifying the conditions of forgiveness, Allais believes she captures what it means to truly "wipe the slate."

Argument

While Allais provides a detailed framework, her assertion that forgiveness requires ceasing personal retributive reactive attitudes makes her account too narrow, failing to reflect the inherent flexibility and variability of forgiveness in real human experiences. Actions necessarily shape how we regard people. Positive actions strengthen our bonds while negative ones weaken them, lowering esteem or trust. This is why forgiveness cannot require the complete affective separation Allais demands. Wrongdoing inevitably reshapes interpersonal relationships, sometimes subtly, or at others, irreparably. Any account of forgiveness that requires victims to return to a previous affective attitude, one in which the wrongdoing no longer figures in their evaluative stance, misunderstands the phenomenology of moral injury. Allais' rigidity stems from her view that personal retributive reactive attitudes must cease for the slate to be wiped clean. She describes these attitudes as complex dispositions that shape one's affective regard for another person. In forgiving, one must forgo this lowered regard. But this overlooks how wrongs leave enduring relational imprints. Human relationships are dynamic and cumulative, past actions inform our expectations, emotional responses, and sense of security going forward. Forgiveness, then, cannot aim at erasing these imprints. Doing so would not only be unrealistic but also could be morally misguided. Sometimes maintaining certain reactive attitudes, like diminished trust or caution, is rational and protective.

Alternatively, I would suggest understanding forgiveness as a spectrum of practices and dispositions rather than a single psychologically uniform achievement. On this spectrum, forgiveness may be as minimal as deliberately overcoming retributive emotions, such as rage or resentment. It may involve a

willingness to rebuild trust, engage in reconciliation, or form new understandings of the wrongdoer's character. Crucially, none of these require an affective reset or the clean removal of negative regard. What they require is that the victim is willing to continue relating to the wrongdoer in some way, whether through renewed cooperation, conversation, compassion, or even just a cessation of hostile attitudes, while still acknowledging the lingering impact of the harm. This broader view is more consistent with our intuitive understanding of forgiveness as something that operates differently depending on the circumstances, relationships, and personal capacities involved. Forgiving a spouse, a stranger, a parent, or a war criminal all involve distinct interpersonal contexts and emotional histories. Allais' account, though, treats forgiveness as structurally identical across contexts. It is always a matter of eliminating personal retributive reactive attitudes and returning to a baseline affective view of the perpetrator. This uniformity erases the lived diversity of moral experience.

Furthermore, Allais' model risks conflating forgiveness with emotional or evaluative forgetting. Victims are not moral archives endowed with the ability, or obligation, to entirely expunge past injuries. Often, they must integrate the wrongdoing into their future understanding of the relationship. In fact, in many healthy post-harm relationships, the wrong becomes a part of a shared narrative of repair rather than something wiped away. For example, a betrayed friend may forgive yet still remain attentive to repeated patterns. A couple may forgive past harms while using those harms as motivation to cultivate better habits of communication. These are typical cases of forgiveness, yet they are incompatible with Allais' requirement of erasing evaluative traces of wrongdoing.

Additionally, Allais' model risks unintentionally moralizing the emotional lives of victims. If forgiveness requires ceasing certain attitudes, then victims who retain lingering wariness, sadness, or disappointment, however rationally and moderately, would count as unforgiving, even when they have wholeheartedly chosen reconciliation. Under Allais' model, forgiveness has specific requirements

that can either be successfully or unsuccessfully met. If not met, then a person who believes they have forgiven will be in a state that is not actually forgiveness. It becomes a concern that treating certain lingering emotions as incompatible with forgiveness portrays victims who forgive imperfectly as a failing, rather than as responding proportionately to harm. Overcoming personal retributive reactive attitudes becomes more about what is owed to the person who did harm rather than the victim's experience and healing process post-harm. Forgiveness, though, must remain elective and cannot demand emotional transformations that may be unavailable or even undesirable. A flexible model instead allows forgiveness to be something victims can offer while still carrying the emotional truths of their experience.

Thus, a richer, more realistic account sees forgiveness not as a slate-cleaning mechanism but as a practice of relational navigation, a way of moving forward while still acknowledging what has occurred. This better captures the moral landscape in which victims operate, where emotional and relational repair unfold gradually, unevenly, and often, incompletely, yet still meaningfully. Such a flexible account may risk collapsing forgiveness into nearby but distinct responses to wrongdoing, such as toleration, resignation, or mere coexistence. If forgiveness is compatible with lowered esteem, diminished trust, and lingering negative reactive attitudes, then it can be called into question what distinguishes it from simply deciding to carry on despite injury. On Allais' view, forgiveness involves a distinctive affective achievement in ceasing to hold the wrongdoing against the wrongdoer. If the wrongdoing continues to shape the victim's affective orientation, forgiveness may seem conceptually emptied of its moral significance.

This highlights that forgiveness must involve a meaningful affective shift if it is to remain a genuinely moral and interpersonal phenomenon rather than a purely strategic or behavioural decision. Acknowledging this, though, does not require accepting Allais' claim that forgiveness consists in the complete cessation of personal retributive attitudes. It is wrong to assume that affective change must be

total in order to be meaningful. Human emotional life is not binary, and this becomes most evident in times of moral and emotional disturbance. Forgiveness does not need to involve an affective reset back to the baseline prior to the wrongdoing in order to constitute a genuine transformation.A flexible account of forgiveness is not distinguished from toleration and resignation by the absence of all negative affect, but rather by a change in the role that those affects play in the victim's moral orientation toward the wrongdoer. Tolerating a wrongdoer would involve continuing a relationship while also continuing to hold negative attitudes that are action-guiding and expressive of ongoing moral protest. Resignation, on the other hand, would involve disengagement from the relationship with the wrongdoer consisting of a freezing of affect that abandons the interpersonal demand altogether. Forgiveness, instead, consists in relinquishing resentment and punitive anger as action-guiding states even when other negative attitudes may persist. What changes is the victim's commitment to holding the wrong against the wrongdoer through retributive affect.

In this sense, forgiveness always involves a meaningful affective shift, but that shift is directional rather than exhaustive. It may take the form of abandoning the desire to retaliate, softening hostility, re-engaging relationally, or affirming the wrongdoer's humanity despite the injury. These are genuine affective transformations even if they coexist with lowered trust or altered esteem. Retaining such attitudes does not indicate an ongoing commitment to blame. Instead, it reflects the rational integration of new moral and epistemic information about the wrongdoer. Understanding forgiveness in this way preserves its moral distinctiveness without imposing unrealistic emotional demands on victims. It allows forgiveness to remain elective and non-obligatory while recognizing that wrongs often leave enduring relational traces. Victims are not morally required to erase these traces in order to forgive. Instead, forgiveness consists in a reorientation that permits forward movement without requiring emotional forgetting or affective neutrality. Forward movement can be achieved either through reconciliation, continued interaction, or internal release. This account

better reflects the lived phenomenology of forgiveness and avoids reducing it either to emotional erasure or to mere endurance.

Case Studies

To best explore the flexibility of forgiveness, it is important to look at ways different harms impact forgiveness, taking into account the role of prior relationships. Allais' lying partner example illustrates this kind of dynamic quite well. Through Allais' account of forgiveness, if the partner that was lied to wants to forgive their lying partner, that would entail no longer holding any opinion of them created by the lying action that they performed. This does not seem realistic. Her account fails to account for the significance of the wrongdoing itself. While the harmed partner may have emotionally moved on from the situation, understanding their partner's reasoning and intentions even though their actions were harmful, I fail to believe that the harmed partner will not be wary of their lying partner's words and actions in the future. This is not because they are still punishing their partner by holding their actions against them, but rather because the act of lying and meeting with their previous partner behind their back revealed to the harmed partner something new about a person they believed to know so well. This naturally alters the nature of their relationship, even if only to a small degree.

Allais might argue that forgiveness is only actually given once the wrongdoing partner proves over time that they will no longer act in such a way and the harmed partner truly believes them. The harmed partner, though, may no longer be hung up on the wrongdoing and may genuinely no longer be upset with their partner. It seems difficult to say that, if future actions spark a potential concern for the harmed partner to believe they are being lied to again, then that would mean they have not forgiven the wrongdoer. Instead, perhaps their relationship has changed so that the harmed partner now knows their partner is someone who would perform such actions.

We can explore a similar dynamic with forgiveness to depicted in a segment in the 2014 Argentinian film Wild Tales titled “Till Death Do Us Part.” During their wedding reception, Romina discovers that her newly wedded husband Ariel had an affair with a coworker, whom he invited to the wedding. Romina confronts Ariel during their first dance and, having been met with dismissal and denial, goes into a spiralling frenzy of revenge ending in blood, tears, and humiliation for Ariel and his mistress. Despite such a reaction, he approaches Romina and extends his hand to offer a dance and a final attempt at reconciliation. Romina decides to take his hand, forgiving Ariel and reconciling with him through a passionate display of affection on the wedding cake table. This chaotic forgiveness defies Allais’ model. Romina’s actions stem from justified resentment, but reconciliation occurs without Ariel’s full repentance, instead just the extension of his hand. Does Romina cease retributive attitudes? Realistically, it is unlikely that, in forgiving Ariel, Romina rescinded her personal retributive reactive attitudes completely. Forgiveness here is flexible, overcoming immediate rage to continue marriage, fueled by the push and pull of the passions. Healing would require ongoing positive acts, such as Ariel’s fidelity, rather than an affective reset. The film’s absurdity highlights how forgiveness in intimacy can be messy, not requiring Allais’ clean separation.

Turning to severe harms in distant relationships, Allais’ TRC cases provide two useful scenarios. In the scenario where the widows forgive de Kock, one offers tears for him and a hand, suggesting empathy. Still, the widow’s hope for his change implies the wrongs still shape her view. She sees him as redeemable but lowered by his acts. Forgiveness here is overcoming emotions to enable dialogue, with healing through de Kock’s apology and potential reform rather than an instant slate-wipe. In Mhlauli’s testimony, the family’s desire to forgive hinges on knowing the perpetrator, but the brutality of the stabbing, acid, and dismemberment permanently alters any view of the killers. If forgiveness occurs, it cannot be said that the families of the victims will be able to feel about the killers the same way they feel about others whom they share no relation with. With an action as serious and severe as torture and murder, it is typically understood that the experience of

performing such actions stays with you and changes your character. Similarly, it permanently alters the way you are seen in the social world. For the victim's daughter to forgive her father's murderer, she cannot separate his action from his character, but she can just as well forgive him by accepting the reality and refusing to dwell on his character.

The case involving Mhlauli's testimony points out an important question about forgiveness when the wrongdoer is unreachable, either because they are unknown or they have passed away. This raises concerns about the possibility of forgiveness when it is impossible. The problem here lies in it not being plausible for forgiveness to aim at restoring an affective stance that Allais' argues is expected in her account of forgiveness. Under a flexible account of forgiveness, this problem is averted by understanding that it will take on a different shape than forgiveness in other situations. If the wrongdoer is unreachable in this way, forgiveness can consist of a reorientation of the victim's emotional state that permits forward movement without affective erasure. It is not possible to restore a previous affective stance. Instead, the wrong can be integrated into the moral narrative without letting it dominate one's agency. In this way, the change happens to the victim's orientation toward the past rather than the relationship they had with the wrongdoer. In all four cases, we see forgiveness is able to adapt to harm and relationship. Minor harms allow quicker emotional overcoming while severe harms demand sustained repair. Allais' rigid forgiveness overlooks this spectrum.

Objection and Reply

It might be argued that, rather than describing the reality of forgiveness, Allais' account of forgiveness captures a core aspirational ideal. Forgiveness can instead be viewed as a practice that may indeed be messy, partial, and gradual, often coexisting with lingering mistrust, sadness, and lowered esteem while still accepting Allais' account of forgiveness, not as a way to describe every ordinary instance of forgiving, but a way to articulate what full or complete forgiveness

would amount to in its ideal form. On this view, the cessation of personal retributive reactive attitudes functions as a regulative ideal rather than a psychological demand placed on victims of harm. This ideal can be a state victims aim at when they want to truly forgive. In this way, Allais might claim that her account preserves the moral meaning of forgiveness without denying the complexity of human emotional life.

This reframing initially appears to soften the rigidity of Allais' position, but it does not ultimately resolve the central difficulties. Even understood as an aspirational ideal, Allais' model continues to privilege a particular emotional endpoint as superior, namely affective neutrality or restored esteem. In doing so, it implicitly evaluates other forms of forgiveness as incomplete or defective, rather than as distinct but equally legitimate ways of responding to wrongdoing. This risks reintroducing, at the level of idealization, the same moralization of victims' emotional lives that is present in her original framework. Victims who forgive while retaining rational wariness or altered regard are still positioned as falling short of the ideal, even when those attitudes are proportionate to the harm suffered.

The aspirational reading also presupposes that what victims ultimately want when they want to forgive is precisely the kind of affective restoration Allais describes. This assumption, though, is not universal. In many cases, what is sought through forgiveness is not the erasure of the wrongdoing's impact on how the wrongdoer is seen, but the possibility of continuing to relate without ongoing hostility, moral stalemate, or emotional paralysis. Victims may aspire to peace, closure, or moral release without aspiring to see the wrongdoer as if the wrong had never occurred. To frame forgiveness' ideal endpoint in terms of affective symmetry or slate-cleaning risks mistaking one moral aspiration among others for the essence of forgiveness itself. It is doubtable that this proposed ideal is normatively attractive. For severe harms, especially those that reveal something enduring about a wrongdoer's character or capacities, a complete cessation of personal retributive reactive attitudes may be neither possible nor desirable. Retaining a modified

affective stance that is shaped by memory, vigilance, or grief can be an expression of moral clarity rather than moral failure. An ideal of forgiveness that requires the victim to bracket these responses risks asking them to adopt an evaluative stance that is insensitive to the moral reality of the harm.

The flexible account does not deny that forgiveness can involve aspirations or ideals. It simply resists the idea that these ideals converge on a single emotional endpoint. Forgiveness may aspire to reduced hostility, restored communication, compassion, or the capacity to move forward without being governed by resentment. These aspirations can be realized in multiple affective configurations, depending on the nature of the wrong and the relationship involved. This pluralism preserves the aspirational dimension of forgiveness while avoiding the imposition of a singular emotional ideal. It is in this way that, even when Allais' view is interpreted as articulating an aspirational ideal rather than a strict criterion, it remains too narrow to capture the full moral significance of forgiveness. Forgiveness does not need to aim at wiping the slate clean, even ideally. It can instead aim at learning how to live with the slate honestly, without allowing it to dictate the future.

Conclusion

Lucy Allas' account of forgiveness succeeds in capturing a powerful moral aspiration in the hope that we can relate to wrongdoers without allowing their faults to define our affective view of them. By insisting that forgiveness must involve the cessation of personal retributive reactive attitudes, however, she constructs a model too rigid for the complexity of real human lives. The case studies, from intimate betrayals to the atrocities addressed in the TRC, demonstrate that forgiveness frequently occurs without an affective reset and often coexist with ongoing emotions and attitudes such as wariness, grief, or lowered esteem. Rather than a singular accomplishment marked by a clean, wiped slate, forgiveness is better understood as a flexible and context-dependent

process through which victims decide how to move forward while still acknowledging the moral reality of what was done to them. This flexible model preserves the elective nature of forgiveness, respects victims' emotional integrity, and accommodates the full range of human relationships.

References

Allais, Lucy. (2008). Wiping the Slate Clean: The Heart of Forgiveness. Philosophy & Public Affairs, vol. 36, pp. 33-68. https://doi.org/10.1111/j.1088-4963.2008.00123.x

Wild Tales. Directed by Damián Szifron, Warner Brothers Pictures, 2014.

END CREDITS

ARTICLE FORMATTING:

blake robertson

GRAPHIC DESIGN:

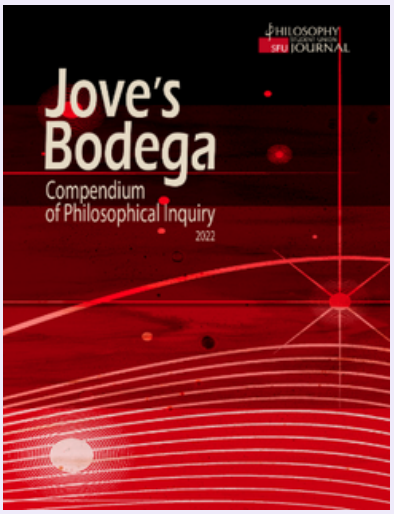
lucia pistrin

MARKETING

paige sangha & ramanjit sahota

PRINTED BY:

sfu doc solutions



Want to see your work in our next issue?

Scan the QR code to access the submission portal.