# *Aporia*

XXIV

# *Aporia*

Undergraduate Journal of the St Andrews Philosophy Society

VOLUME XXIV

Aporia is funded by the University of St Andrews Philosophy Society, which receives funds from the University of St Andrews Department of Philosophy, the University of St Andrews Students' Association, and independent benefactors.

# Letter from the Editor

Dear Reader,

As we close another academic year, I am delighted to present the latest edition of *Aporia*. First and foremost, I would like to extend my gratitude to the remarkable team of editors who have worked tirelessly alongside me. Their dedication and hard work have made this publication possible. Special thanks are owed to Deputy Editor-in-Chief, Kirsty Graham, and Technical Officer, Mohit Agarwal, for their unwavering commitment to the journal, and whose efforts ensured the creation of a publication that I hope we can all be proud of.

This year, I also had the privilege of collaborating closely with Christina Landys-Herre, the editor of the Feminist Edition of *Aporia*. Her passion and enthusiasm were a constant source of inspiration, and her invaluable support was indispensable in bringing this edition to life.

The current volume offers valuable insight into the state of undergraduate academia in the realm of analytic philosophy today. It reflects both the diversity and the unity of the discipline we all hold dear, and I hope that it will inspire philosophers-to-be as they explore its pages.

Finally, a heartfelt thank you to the University of St Andrews Department of Philosophy, as well as to everyone who submitted their work this year.

I truly hope you enjoy this edition.

Yours faithfully,

Avery Cohen

# Acknowledgements

Avery Cohen
*Editor-in-Chief*

Kirsty Graham
*Deputy Editor*

## Editors

Mohit Agarwal

Joe Bradstreet

Alastair Bowyer

Olivia Griffin

Anders Knospe

Yinan Li

Faye Perry

Nathalie Rogers

James Shearer

Lahari Thati

Lara Thain

Rosa Velasco

Eric Wallace

Finn Wheatley

Hoochang Yi

Mohit Agarwal
*Technical Officer*
*Cover Art*

# Contents

# What's the 'Puzzle about Belief'? Revisiting Kripke's Challenge through a Fregean Lens

Ethan Reiter, *University of Chicago*

In "A Puzzle About Belief" (1979), Saul Kripke posits a thought experiment involving a bilingual named Pierre, who comes to what seem like contradictory beliefs about the city of London. On a popular reading, the upshot of how Kripke sets up Pierre's problem is that the challenges of ascribing propositional attitudes such as belief to another person are more complex than suggested in previous philosophical analyses, particularly Gottlob Frege's "Sense and Reference" (1948). In particular, Kripke aims to show that a genuine paradox about belief ascription can arise without the need to presuppose that a thinker with different beliefs about the same object relates to that object through different cognitive perspectives (i.e., what Frege called modes of presentation). In this essay, I take issue with how Kripke draws his conclusion that Pierre's problem must go beyond Frege's notions of sense and reference. I argue that application of Fregean notions of sense indeed 'solves' Kripke's puzzle about Pierre and that Kripke's pre-emptive reply to such an objection does not suffice, casting doubt only on how the Pierre paradox gets off the ground. In sum, I argue that a charitable understanding of Frege's ideas and a closer interrogation of Kripke's set-up reveals that Pierre's 'puzzle' is not necessarily a puzzle.

As the trajectory of analytic philosophy makes clear, describing the thought and belief of others with language is no trivial feat. Often, situations in which Leibniz' law regarding identity[1] apparently fails to hold arise. In "Sense and Reference" (1948), Gottlob Frege highlights these logical problems encountered in situations of thought attribution. He argues that the reason Leibniz' law ostensibly breaks down is that when attributing thoughts to thinkers, we must also attribute their occupying a particular cognitive stance on the objects of their thought. Frege thus introduces the sense-reference distinction to show how intersubstitution of coreferential simple proper names in identity statements and ascriptions of apparently contradictory beliefs to an agent need not jeopardize Leibniz' law. Meanwhile, in "A Puzzle About Belief" (1979), Saul Kripke argues that invoking cognitive perspectives in belief is beside the point. He attempts to show that the same intersubstitution failures can be generated with the disquotation and translation principles alone,[2] not requiring any sort of belief perspective. Kripke puts forward the Pierre puzzle, one version of which supposes that even when a subject adopts the same cognitive perspective on two different but coreferential signs, contradictory beliefs can still arise. This paper reconstructs and juxtaposes Frege and Kripke's puzzles to demonstrate the import of Kripke's critique. However, the paper concludes that Kripke's puzzle does not necessarily amount to a puzzle, given that the Fregean line has its own resources to make sense of Pierre's situation through reiterating the extent to which cognitive perspectives shape thought.

[1] I refer here to Leibniz' statement that "those things are the same of which one can be substituted for another without loss of truth," a principle which Frege also committed himself to.

[2] While the debate over these principles is interesting, it is not the topic of this paper. Thus, this paper will limit itself to critique of Kripke's set-up of his puzzle rather than the principles it presupposes.

## 1 Preliminary Remarks about Frege and Kripke's Puzzles

For Frege, contexts in which Leibniz's law of identity produces unacceptable results can be resolved through considering thinkers' cognitive stances on their objects of thought. Frege's first puzzle contrasts equality statements like 'a = a' and 'a = b' (provided that 'a' and 'b' designate the same object), stating that while the truth of 'a = a' is trivial and "*a priori,*" "statements of the form *a = b* often contain very valuable extensions of our knowledge and cannot always be established *a priori*" (Frege 1948, 209).[3] That is, in the latter construction, one simply cannot intersubstitute 'a' and 'b' without altering its "cognitive value," despite their coreference (209). Frege's second puzzle, regarding ascription of propositional attitudes like belief, is even more illustrative: statements can differ not only in their informativeness, but also in their *truth value*, even when all which distinguishes them is the substitution of coreferential names. For example, if Venus is in orbit, one might say upon waking up, "I believe I will see the morning star right now" *and* "I do not believe I will see the evening star right now": insofar as one (naturally) resists ascribing irrationality to such an agent for having contradictory beliefs, those coreferential designators of Venus must convey different meanings, such that sense is one thing, and reference another. For Frege, the 'morning star' and 'evening star' represent the different—-and limited—-cognitive perspectives thinkers can take on objects, as seen easily in such cases of discovery, as well as confusion.[4] These distinct "mode[s] of presentation" (210), as Frege calls them, differ in their cognitive value[5]—-"the thought changes," therefore, when intersubstituting distinct senses (215). Consequently, Frege argues that both his first puzzle on identity statements and his second puzzle on propositional attitude ascription involve a thinker taking two irreducibly different cognitive stances on (i.e., senses of) the object of their thought.

Here, it is worth bringing Kripke's views on such cognitive perspectives into focus: he argues that considering sense distinctions is not necessary to create the sort of intersubstitution problems which troubled Frege. Instead, Kripke only presumes the disquotation principle and the translation principle.[6] He poses a problem, by appeal to the imaginary example of Pierre, of a bilingual speaker who holds what seem to be contradictory beliefs about the same proper name, solely by virtue of holding those beliefs separately in his two languages. Pierre grows up a standard French speaker, believing that London is pretty. Later in life, he becomes a standard English speaker by direct method[7] after unknowingly ending up in London (finding, then, that it is *not* pretty). He thus holds the following two beliefs, according to the disquotation principle:

(a) in French: '*Londres est jolie*'

(b) in English: 'London is not pretty'

[3] Frege provides the example of the referent Venus, which is designated by both the names 'the morning star' and 'the evening star.' While 'the morning star is the morning star' strikes one as empty-headed, Frege writes that 'the morning star is the evening star' "was of very great consequence to astronomy. Even today the identification… is not always a matter of course" (209).

[4] A classical example is how readers understand the tragedy of Oedipus Rex: as thinkers like Jerry Fodor have noted, one can attribute to Oedipus the belief that he wants to marry Jocosta but *not* the belief that he wants to marry his own mother, and it is by virtue of this discrepancy that the text is ultimately understood as a tragedy.

[5] Put another way, the names being intersubstituted do not have the same psychological value to their thinker: they are distinct perspectives on, different senses of, the object of thought (i.e., Venus). Thus, the sense-reference distinction is what empowers 'a = b' to be informative, whereas 'a = a' is not, in that it conveys *new* knowledge that two senses are indeed of the same referent.

[6] For the sake of my argument that Kripke's puzzle can be resolved through a Fregean lens, these two principles (in their various weak and strong forms) can be taken for granted and set aside. To briefly summarize them, following Kripke: the disquotation principle holds that when a speaker sincerely assents to a proposition, they believe it (439); the translation principle holds that the truth-value of a sentence in one language is preserved upon translation to another (440).

[7] That is, Pierre learns English entirely without the help of his native language, French. Rather than learning English names through translation from their French equivalents, for example, Pierre learns English as though it is his first language, connecting English expressions directly to his audiovisual experiences of the world. Consequently, he fails to realize '*Londres*' is 'London.'

These beliefs form an outright contradiction, at least if the translation principle is applied to (a) to result in 'London is pretty' and Kripke's ascription of both beliefs to Pierre is accepted.[8]

If both Pierre's French and English beliefs about London are to be upheld, though, then how can Pierre avoid contradictory beliefs? Indeed, Kripke clarifies that "as long as he is unaware that the cities he calls 'London' and '*Londres*' are one and the same, [he] is in no position to see, by logic alone, that at least one of his beliefs must be false" (444). Even if Pierre is a logician par excellence, he would be unable to leverage *modus tollens* to correct his 'contradictory' beliefs until he realizes his beliefs about 'London' and '*Londres*' actually regard the same city. This gives rise to Kripke's thesis: "that the [Pierre] puzzle *is* a puzzle" (433). It is difficult—-paradoxical even, Kripke says—-to state Pierre's true beliefs about the city of London.

Could this Pierre puzzle resemble Frege's second puzzle about belief ascription, inviting clarification that 'London' and '*Londres*' are distinct cognitive perspectives on (i.e., senses of) the city of London? At first glance, Kripke admits, "[o]ne aspect of the presentation may misleadingly suggest the applicability of Frege-Russellian ideas that each speaker associates his own description or properties to each name" (445). However, Kripke swiftly dismisses that '*Londres*' and 'London' could be considered distinct modes of presentation of London in that it is not required to assume that the two translations of the city of London's name satisfy distinct sets of properties. Indeed, Kripke argues that "the puzzle can still arise even if Pierre associates to *'Londres'* and to 'London' *exactly* the same *uniquely identifying* properties," such as in the case that he identifies both as having the very same landmarks, monarchs, etc., insofar as he "regard[s] *both* properties as uniquely identifying" (446). Since Pierre acquires English through the direct method, Kripke maintains that nothing compels Pierre to equate terms like '*Angleterre'* with 'England'; Kripke avoids presupposing "an 'ultimate' level [of language]… where the defining properties are 'pure' properties not involving proper names" (447). Kripke thus concludes that even a shared set of uniquely identifying properties for 'London' and '*Londres*' could not alone compel Pierre to infer any contradiction from his separately held beliefs. Given such perils, Kripke concludes that Pierre's paradox goes deeper than Fregean notions of sense.

## 2 The Potential for a Fregean Reinterpretation of the Pierre Puzzle

I will argue that Kripke's swift dismissal of Fregean readings does not withstand scrutiny. Specifically, Kripke's rejection that "Pierre believes that *the city* satisfying *one* set of properties is pretty, while he believes that *the city* satisfying *another* set of properties is not pretty" (445) ought to be reconsidered. Kripke was too hasty to assume that Frege's invocation of cognitive perspectives has no place within the puzzle he sets up. Indeed, Pierre's contradictory beliefs, particularly about the prettiness of London, can hardly be accounted for in the first place if Pierre is said to truly occupy the same cognitive perspective in both his English and French belief sets.

Ultimately, the identifying properties of '*Londres*' associated with Pierre's French belief that the city is pretty are fundamentally different from the properties of 'London' associated with his English belief that the city is not pretty. Recall how Kripke sets up Pierre's puzzle. The set of properties informing Pierre's beliefs about '*Londres*' are provided "[o]n the basis of what he has heard of London" in French—-those positive descriptions which make him "inclined to think that it is pretty" (442). Pierre's belief of '*Londres*' thus corresponds to the compliments that he has heard about the city—-they uniquely identify the city's prettiness for him. Yet there are no such compliments which correspond to 'London' for Pierre. There is likewise no such corresponding

[8] I will follow Kripke in ascribing both beliefs to Pierre. If one tries to undermine Pierre's old belief in French that London is pretty, this entails the absurd conclusion that other monolingual French speakers must lack belief about the prettiness of '*Londres*.' But if one tries to undermine Pierre's new belief in English that London is *not* pretty, they mistakenly think "Pierre's French past [can] nullify such a judgment" (444). Kripke suggests considering "an electric shock [which] wiped out all his memories of the French language, what he learned in France, and his French past": but this, too, creates a *reductio* insofar as Pierre cannot *gain* a new belief from destruction of his memory, and Pierre should hold the same belief as his new English countrymen (444). Combining the two denials of belief "in his bilingual stage" would only compounds these difficulties (444).

property, 'this is reputed by my countrymen as pretty,' for 'London'; in fact, Kripke asks readers to imagine the opposite, that Pierre's English neighbors "rarely venture outside their own ugly section" (445). Given such differences, Pierre occupies a specific cognitive stance on '*Londres.*'

Meanwhile, the set of properties informing Pierre's beliefs about 'London' are provided when Pierre "is unimpressed with most of the rest of what he happens to see" while in England (443). Indeed, when Pierre thinks of 'London,' he likely recalls the filth of the part of the city wherein he now lives. There is no corresponding uniquely identifying property, 'part of this city where I lived is filthy,' which Pierre could recall when thinking of '*Londres*' insofar as (like Kripke points out) he does not even realize he is living in '*Londres*.' This explains how Pierre occupies a particular cognitive stance on 'London,' distinct from the stance which he occupies on '*Londres*.' Especially given that Pierre learns English through the direct method, the set of experiences which Pierre uniquely associates with English expressions like 'London' date much later than the set of experiences which Pierre uniquely associates with French expressions like '*Londres.*' Pierre just cannot be said to have become acquainted with the two names for this city in the same way: for all he knows, he does not even live in '*Londres*.' The upshot is that the set of properties on which Pierre bases his beliefs of 'London' simply are *not* shared for '*Londres*.' Insofar as Pierre never realizes that 'London' and '*Londres*' are the same, Kripke must concede that Pierre is in no position to realize that the reputation of '*Londres*' as pretty uniquely identifies the *same* city as does Pierre's uniquely identifying memory of living in an ugly part of 'London.' The set of properties defining 'London' for Pierre, then, are not those which define '*Londres*.'

Consequently, Kripke cannot be correct when he writes that "the puzzle can still arise even if Pierre associates to *'Londres'* and to 'London' *exactly* the same *uniquely identifying* properties" (446): if this was true, which uniquely identifying property shared between both '*Londres*' and 'London' could give rise to contradictory beliefs about whether the city is pretty? As suggested, this property could neither be the city's reputation nor Pierre's living experiences, given that both correspond to Pierre's *distinctive* cognitive stances on 'London' and '*Londres*.'

There is thus simply no potential for these two subsets of uniquely identifying properties (i.e., those which designate '*Londres*' pretty and those which designate 'London' not pretty) to be isomorphic given their irreconcilable implications for the prettiness of the city of London. At some point, something has to give: some uniquely identifying properties of '*Londres*' make it pretty and *other* uniquely identifying properties of 'London' make it *not* pretty if Pierre truly reaches opposing conclusions about the two names in his beliefs.[9] This Fregean reading makes the straightforward concession that Pierre believes that '*Londres*' is pretty while he believes that 'London' is not, grounding the inconsistency in the names' distinctive modes of presentation. On such a view, Pierre's puzzle is no less tractable than Frege's second puzzle concerning failures intersubstituting coreferring names in belief contexts: Pierre's *ostensibly* conflicting beliefs about London's beauty merely reflect the two irreducibly distinct cognitive stances he occupies of it.[10]

Indeed, consider when Kripke compares Pierre's situation to one who comes to different beliefs about the baldness of 'Plato' (English) and '*Platon*' (French), figuring they are different people (446). Kripke introduces this second scenario with the express purpose of pre-empting the Fregean reading this paper defends, illustrating instead how "[t]he puzzle can arise even if Pierre associates exactly the same identifying properties with both names" (446). However, the Fregean counterpoint still remains: what *shared* properties could be the basis for the baldness attributed to

[9] It may be objected at this stage that even if Pierre's belief sets in English and French are the same, he could nevertheless draw different conclusions about the city of London through the explosion principle of classical logic. For the present purposes, this reply can be set aside. It is not a straightforward matter to generalize from classical logic to the psychology of belief, especially in regard to the explosion principle: what it would mean to truly believe both $p$ and $\sim p$—-and in what sense would that entail a believer to believe whatever else they liked? Moreover, this paper follows Kripke in presuming Pierre is a skilled logician; Pierre cannot be admitted to (knowingly) believe in two inconsistent sets of premises, so he could not make use of explosion anyway.

[10] As this paper will explain, the Fregean can analogize Pierre's beliefs that '*Londres*' is pretty and that 'London' is not to the example provided earlier, of one's belief that they will see the 'morning star' but not the 'evening star' at the beginning of the day. In both cases, there is no true contradiction, insofar as the object of one's belief—-and thus the belief—-changes when the sense changes.

'*Platon,*' contra the hirsuteness attributed to 'Plato'? The property on the basis of which 'Plato' has hair is necessarily different from the one on which '*Platon*' is bald. Just as in the Pierre case, it is incoherent to imagine the two names for the Greek philosopher having "exactly the same identifying properties" (446) while at once differing in the very property that makes them different when they are taken as objects of belief—-the crux of the apparent contradiction. If Kripke means to argue that there is truly no psychological difference between the English and French equivalents in the Plato case (just as in the Pierre case), it is unclear how exactly the contradictory beliefs could surface to begin with. Is that not evidence enough of a difference?

Kripke makes the incongruity between Pierre's French and English belief sets seem to stem from a whim—-as though some distinction between just the *valence* of French versus English descriptions of London could have led Pierre to arrive at a different conclusion about the city's prettiness from shared uniquely identifying properties. This is a critical oversight: how can one say Pierre's starting assumptions for arriving at his beliefs about the prettiness of 'London' and '*Londres*' stem from the exact same sets of information? They obviously do not. At some point in cataloging Pierre's French and English belief sets, there *must* be discord (particularly in those beliefs bearing on judgments about the city's prettiness)—-or else, what could possibly account for Pierre reaching different conclusions about its prettiness in the first place? Unless this divergence is recognized for what it is, Kripke would need to make an entirely different argument to contextualize how Pierre arrives at contradictory beliefs. If 'London' and '*Londres*' truly do share all the same uniquely identifying properties, Kripke must pass the explanatory buck for Pierre's contradictory beliefs about the two to intrinsic differences between French and English. After all, only those could plausibly dispose Pierre to judge one thing about the former yet the opposite about the latter if he truly assumes the same cognitive stance toward them both.

In this way, the discrepancy between Pierre's French and English beliefs about London's prettiness is akin to Frege's puzzle, wherein one believes they are seeing the 'morning star,' but not the 'evening star,' upon waking up. Just as how in the latter, the difference between '*a*' and '*b*' is a uniquely identifying property which distinguishes the two modes of presentation (e.g., based on the time of day when Venus is in view or the stage of its orbit), the difference between 'London' and '*Londres*' is those uniquely identifying properties which make the former ugly (e.g., the hideosity of Pierre's neighborhood) but the latter pretty (e.g., the city's reputation of beauty among French speakers). Appreciating this difference further allows for recognizing that Kripke's ultimate question—-"Does Pierre or does he not, believe that London (not the city satisfying such-and-such description, but *London*) is pretty"—-is falsely posed (446). Since Pierre occupies two distinctive cognitive stances on "*London*" in his beliefs, Pierre's beliefs about the city cannot but be understood in Fregean terms of modes of presentation. In the final analysis, Pierre's beliefs of 'London' are irreducibly different from his beliefs of '*Londres*,' as with the 'morning star' and 'evening star.' If Kripke argues that 'London' and '*Londres*' are not two different senses of London to Pierre, he need not only refute the above consideration, but also meet the explanatory burden of *how* Pierre could ever come to contradictory beliefs about the city's prettiness if, in both his languages, the set of uniquely identifying principles are isomorphic. Ultimately, a Fregean can reintroduce cognitive perspectives into Pierre's puzzle, casting doubt on its paradoxicality through concluding that Pierre's beliefs concern different modes of presentation of the city of London after all. Indeed, they stem from different thoughts.

## 3 Conclusion

This paper suggests a Fregean resolution to the ostensible paradox about Pierre which Saul Kripke puts forward in "A Puzzle about Belief" (1979). I argue that for Pierre, 'London' and '*Londres*' are two distinct modes of presentations, two distinct senses, of the city of London. When making sense of what Pierre believes about the English capital, 'London' and '*Londres*' need not be taken to connect to the same thought: I argue that for Pierre, the names *do* correspond to cities with two *necessarily* distinct sets of uniquely identifying properties. Indeed, the necessary distinction is that

on the basis of which Pierre draws opposite conclusions about whether London is pretty—-the same distinction that engenders his key 'contradiction' in belief.

In Kripke's set-up of the puzzle, Pierre comes to contradictory beliefs about London's prettiness based on an incongruity between what he hears about '*Londres*' from his countrymen while living in France and what he sees on the ground while living in 'London.' I argue that Kripke is wrong to think that both opposing aspects of this inconsistency can be accommodated within the same cognitive stance on the part of Pierre if Pierre never realizes that 'London' and '*Londres*' are coreferring names. Instead, Frege's sense-reference distinction remains crucial: Pierre's puzzle is not paradoxical because he occupies two distinct cognitive stances on the city of London, such that his incongruous beliefs about 'London' and '*Londres*' do not contradict each other—-they stem from altogether different thoughts. Thus, Kripke's 'puzzle' about belief does not amount to a puzzle, so long as Pierre's cognitive perspectives in his beliefs about 'London' and '*Londres*' are distinguished along the line Frege endorses with his second puzzle.

## Bibliography

Frege, Gottlob. "Sense and Reference." *The Philosophical Review* 57, no. 3 (May 1948): 209–30. https://doi.org/10.2307/2181485.

Kripke, Saul A. "A Puzzle about Belief." *Meaning and Use*, 1979, 239–88. https://doi.org/10.1007/978-1-4020-4104-4_13.

# Intellect, Personhood, and the Coherence of Cure

Ethan Samuel Kovnat, *Cornell University*

People tend to associate intelligence with personhood, and likewise, there is a tendency to tacitly deny personhood to those with intellectual disabilities. However, I argue that this denial of personhood is not automatic, and I will draw from recent work in the social psychology of autism to explain the role of empathy in ascriptions of personhood. I will argue that the tendency to deny personhood to people with cognitive and intellectual disabilities is the result of a lack of reciprocal empathy. I will then describe how this is illustrative of key differences between how physical disabilities and cognitive/intellectual disabilities impact one's identity and explain how these differences should inform our understanding of whether the idea of curing cognitive/intellectual disabilities is coherent.

In his book *Brilliant Imperfection: Grappling with Cure*, Eli Clare criticizes the longstanding association between intelligence and personhood. According to Clare, it is a mistake for intelligent disabled people to assert their personhood on the basis of their intelligence, as this perpetuates the marginalization of and denial of personhood to the intellectually disabled. I find Clare's argument partially convincing; while I think Clare is right in his observation that the intellectually disabled as not treated as people of the same status as the nondisabled, and while I agree with his larger point that using intelligence to justify personhood is deeply problematic, I consider Clare's description of how people are dehumanized for their lack of intelligence to be an oversimplified account of how intellectually disabled individuals are actually denied equal personhood. I do not think society treats lack of intelligence as itself grounds for dehumanization, but rather, that the difficulty for non-intellectually disabled people to grasp what it is like to be intellectually disabled leads to a broad failure to understand that intellectually disabled people, and others with neurodevelopmental conditions, are people in the same way that the nondisabled are. Understanding the reasons behind this will reveal an important difference in the modalities of intellectual/neurodevelopmental disabilities and physical disabilities. I argue that this distinction has broad ramifications, and should particularly inform our politics of cure.

The link between intellect, specifically rational intellect, and personhood is deeply ingrained, both in philosophical and popular discourse. In the philosophical literature, this idea can be traced back at least as far as Boethius, who defined a person as "an individual substance of a rational nature."[1] Similar accounts can be found in the works of Descartes, Locke, and Hume, all of whom took the ability to formulate rational plans of action to be a core element of personhood.[2] However, the idea that rationality as a necessary condition for personhood was perhaps most influentially articulated by Kant:

> …every rational being, exists as an end in himself and not merely as a means to be arbitrarily used by this or that will…Beings whose existence depends not on our will but on nature have, nevertheless, if they are not rational beings, only a relative value as means and are therefore called things. On the other hand, rational beings are called persons inasmuch as their nature already marks them out as ends in themselves…[3]

[1] Joseph W. Koterski, "Boethius and the Theological Origins of the Concept of Person," *American Catholic Philosophical Quarterly* 78, no. 2 (2004): 203–24, https://doi.org/10.5840/acpq200478212.

[2] Charles Taylor, *The Concept of a Person* (Cambridge, UK: Cambridge University Press, 1985), 97-114

[3] Immanuel Kant, *Groundwork for the Metaphysics of Morals*, trans. Mary J. Gregor (Cambridge, UK: Cambridge University Press, [1785] 1998), 4:428. Quoted in Lori Gruen, "The Moral Status of Animals," Stanford Encyclopedia of Philosophy, June 23, 2021, https://plato.stanford.edu/entries/moral-animal/.

Kant's definition of personhood is particularly important because it directly ties rationality, and therefore personhood, with moral value. To Kant, a person is defined as something that is an end in itself, i.e. deserving of moral consideration for one's own sake,[4] and in order to be an end in oneself, one must possess a rational intellect.

This concept of personhood seems to have made its way into social ascriptions of personhood. Although it is not one of the central topics discussed in *Brilliant Imperfection*, Clare makes a point to set aside a brief section in which to discuss the relationship between intelligence and ascriptions of personhood. Clare's thesis here is that because "intelligence is used repeatedly to determine worthiness, value, and personhood,"[5] disabled people who are not intellectually disabled are often compelled to justify their personhood by asserting their intelligence. According to Clare, disabled people, along with racial minorities and queer people, are stereotyped as being "defective" or "intellectually inferior," and are often forced to use their intelligence to fight against these perceptions. Clare, who is physically disabled, cites examples of this phenomenon from his everyday life: "I think about the ways I defend myself when the bullies call *retard* and grocery store clerks, doctors, teachers, or strangers on the street talk to me loudly and slowly as if I can't understand. My most immediate response is to declare myself smart, not intellectually disabled… I've repeatedly used intelligence as the marker of my worth and personhood."[6]

This social reality reveals a straightforward tension in any definition of personhood, either as a metaphysical status or as a social role, in which a person is defined by their intelligence: Inasmuch as a person is a being capable of rational intellect, an intellectually disabled individual cannot be a person. Although providing a philosophical account of personhood in general is far beyond the scope of this paper, I do want to take two statements about personhood as given. Firstly, I want to take it for granted that one quality closely associated with personhood is consciousness, or, as Thomas Nagel would put it, that there is "something that it is like to *be*" a person.[7] While I am not prepared to claim categorically that consciousness is a necessary condition for personhood, I think it is fairly uncontroversial to suppose that whether an entity possesses a conscious experience plays a substantial role in whether we give them the moral consideration of a person. Secondly, I want to take it for granted that any viable account of personhood cannot exclude the intellectually disabled. This is perhaps a slightly more controversial point,[8] though few would deny that it is an intuitive one, and it is certainly one that Clare accepts.

As such, Clare notes that when people assert their intelligence to signal that they are worthy of personhood, they tacitly suggest that lack of intelligence is grounds for denial of personhood, contributing to the marginalization and mistreatment of those who are intellectually disabled. "Every time we defend our intelligence," Clare argues, "we come close to disowning intellectually disabled people. We imply that it might be okay to exclude, devalue, and institutionalize people who actually live with body-mind conditions that impact the ways they think, understand, and process information."[9] As such, Clare sees it as imperative that we "resist using intelligence as a measure of worth and personhood"[10] so as to fight against these horrific attitudes toward the intellectually disabled.

Clare makes two claims in this argument: one descriptive and one prescriptive. The descriptive claim is that society uses intelligence as a marker of personhood, and that this denies personhood to the intellectually disabled. The prescriptive claim is that we must resist this tendency so as to defend intellectually disabled people. I want to address the prescriptive claim first, because regardless of the truth of the descriptive claim, the moral imperative Clare outlines seems impossible to deny.

[4] Kant, *Groundwork for the Metaphysics of Morals*, 4:428

[5] Eli Clare, *Brilliant Imperfection: Grappling with Cure*, (Durham: Duke University Press, 2017), 156

[6] Clare, *Brilliant Imperfection*, 157

[7] Thomas Nagel, "What Is It Like to Be a Bat?," *The Philosophical Review* 83, no. 4 (October 1974): 435, https://doi.org/10.2307/2183914.

[8] Some characterizations of Peter Singer's views of personhood and moral status maintain that he excludes the intellectually disabled from either or both. However, it is unclear to me if such readings of Singer accurately reflect his position, and I fear that some of these are made in bad faith.

[9] Clare, *Brilliant Imperfection*, 158

[10] Clare, *Brilliant Imperfection*, 158

Intellectually disabled people deserve to be ascribed the same personhood as any other human being, and any ideology that denies them that personhood is would that we must certainly resist. So, even if we are inclined (which I am, to some degree) to dispute Clare's descriptive claim that lack of intelligence is itself actually taken as grounds for dehumanization, that does not change the fact that we cannot use intelligence as a justification of personhood without casting doubt on the status of the intellectually disabled as people. As such, I agree wholeheartedly with the broader point of Clare's argument that it is exclusionary for anyone to use intelligence to justify their own personhood, even if it may seem necessary or expedient to do so in our social environment.

It is Clare's descriptive claim that I would like to analyze more closely. Clare seems to treat lack of intelligence as itself the basis for dehumanizing treatment, but he does not go into detail about the mechanism by which this dehumanization happens. Rather, per my reading of Clare's text, he would seem to take this dehumanization to be automatic: in keeping with Kantian-style accounts of personhood, society recognizes lack of intelligence and then interprets it as lack of personhood and moral value. Regardless of whether or not this reading is what Clare intends, I want to react against the notion that lack of intelligence is alone and automatically the root of intellectually disabled people being denied their social personhood. It seems to me that this denial is a more complicated process, and that there are steps between recognizing intellectual disability (or neurodevelopmental disability, for that matter) and dehumanization. To explain this process, I want to move slightly away from discussing intellectual disability, and shift my focus to autism, because there is relevant literature dealing with autism specifically, and the case of autism lends itself to fruitful practical examples. While autism and intellectual disability are frequently comorbid, many autistic people are not intellectually disabled, and discussing some of the ways in which non-intellectually disabled autistic people are viewed by society can be helpful in understanding the social status of intellectually/neurodevelopmentally disabled people in general.

I want to begin by explaining the double empathy problem, which was first formalized by social psychologist Damian Milton, who is himself autistic. The double empathy problem was presented as an argument against the conventional view that autistic people lack a theory of mind, and that difficulty socializing is therefore inherent to autism. Rather, Milton takes the social challenges characteristic of autism to be the result of a lack of reciprocity of empathy between autistic and allistic (that is to say, non-autistic) interlocutors. According to Milton, because allistic people cannot identify with the lived experiences of autistic people, allistic people are not able to "assume understandings of the mental states and motives" of their autistic peers with the same reliability with which they are able to understand their allistic peers.[11] Meanwhile, because autistic people typically experience difficulty interpreting other people in general, and thus have a similar difficulty understanding their allistic peers. In both cases, some factor gets in the way of cognitive empathy, and this mutual lack of understanding, or failure of double empathy, explains why autistic people often find social interaction (with allistic people in particular) to be so perplexing.[12]

While the double empathy problem is intended to be an explanation for why autistic people typically struggle socially, I think it also illustrates why and how autistic people, as well as others with intellectual/neurodevelopmental disabilities, are so often denied personhood. Because allistic people cannot reliably understand the inner lives autistic people to the same degree that they can understand the inner lives of other allistic people, it seems plausible that it is easy for many allistic people (absent proper reflection) to lose sight of the fact that autistic people have inner lives at all. When the existence of something is not readily apparent, it is natural to disregard its existence altogether; it is natural for an allistic person to proceed as though there is not "something that it is like to *be*" an autistic person. And because autistic people typically struggle to understand their allistic peers in the same way that their allistic peers struggle to understand them, autistic people often find themselves unable to display that they have inner lives in a way that is perceptible to

[11] Damian E. M. Milton, "On the ontological status of autism: the 'double empathy problem,' " *Disability & Society* 27, no. 6 (2012) 883-887, accessed October 30, 2023 https://www.tandfonline.com/doi/full/10.1080/09687599.2012.710008.

[12] Milton, "On the ontological status of autism"

allistic people. The point here is easy for an allistic person to proceed as though autistic people lack the conscious experience that is so closely tied to personhood.

Of course, at least in the case of most allistic people, the assumption that autistic people lack inner lives is not a conscious one; surely, on an intellectual level, most allistic people do recognize that autistic people do have conscious inner lives. But this recognition, no matter how genuine, often fails to penetrate the popular discourse surrounding autism. For one thing, many of the ways in which autism is pathologized seem to deny that there is conscious experience behind the actions of autistic people. For an example of this, see the page of the Autism Speaks website titled "What are the Symptoms of Autism?" This page describes various traits common among autistic people; most illuminating is the section in which it describes "Restricted and repetitive behaviors" where it states that autistic people, among other things, tend to have "repetitive body motions (e.g. rocking, flapping, spinning, running back and forth," tend to display "repetitive motions with objects (e.g. spinning wheels, stacking sticks, flipping levers" and tend to engage in "ritualistic behaviors (e.g. lining up objects, repeatedly touching objects in a set order)."[13]

What is absent from these descriptions, though, is any acknowledgment that these repetitive or ritualistic actions have a conscious experience associated with them. There is no acknowledgment that repetitive motions, what neurodivergent people like to call "stimming," helps relieve anxiety and bring calm in the face of overwhelming sensory experience. Nor is there acknowledgment that ritualistic behaviors bring a sense of structure and security, and that dispensing with them can be incredibly disturbing. When autism is pathologized in this way, typical autistic traits are treated as the automatic actions of a robot running some sort of autism program. There is no recognition of the conscious lived experiences that actually define who an autistic person is. The actions of an autistic person are not treated as the actions of a person, but as actions of autism.

For another thing, there is a tendency to reduce autistic people's successes to a function of autism itself, not to individual skill or perseverance. The ubiquitous slogans of, "Autism is not a disability, it is a different ability!" and "Autism is your superpower!" are just symptoms of the widespread attitude that when autistic people are successful, it is because they are autistic. Personally, I have lost count of the number of times I have been praised for my intelligence, and then told that I must be intelligent simply because I am autistic, since "autistic people have bigger brains." Successful autistic people are quickly labeled autistic savants, and media representations of autism tend to highlight the "mildly autistic super-detective,"[14] or the autistic doctor or computer hacker who can instantly understand and visualize complex systems simply by activating their magic autism powers.[15] Once again, the human experience of what it is like to be autistic is ignored; autistic people are only seen as valuable when they perform function.

And, finally, it is so very rare to see public discourse about autism in which autistic people are actually involved. So much of this space is occupied by caregivers of autistic people,[16] and oftentimes, the opinions being expressed do not accurately represent those found among the autistic community. Usually the excuse for this is that autistic people lack the competence or communication skills necessary to effectively self-advocate, and so their caregivers need to advocate for them.[17] While it may be true that some autistic people, particularly those with accompanying intellectual disabilities, may not be able to self-advocate, this is certainly not true of all autistic people, as evidenced by the existence of organizations like the Autistic Self-Advocacy Network. And yet, it is still mostly allistic people who manage to break into the conversation. It seems to me that the relative absence of autistic people from this public discourse stems from the idea that the legitimate people worth hearing from must not be autistic, and that autistic people are not

[13] "What Are the Symptoms of Autism?" Autism Speaks, accessed October 30, 2023 https://www.autismspeaks.org/what-are-symptoms-autism.

[14] A term borrowed from the television series *Community* (NBC, 2009-2015), which satirizes this trope.

[15] For a good example of this, see the television series *The Good Doctor* (ABC, 2017-present). Watch at your own risk.

[16] Often, such individuals who enter the public discourse are referred to as "autism moms." I prefer not to use this term, as I do not want to make unfair generalizations about the mothers of autistic people (my mother, for instance, is absolutely wonderful), but this is the phenomenon I am referring to.

[17] This is the ideology on which Autism Speaks is built.

legitimately people in the same way. I think all three of these examples show the same things: a) many allistic people are not able to empathize with autistic people, which leads to b) these allistic people failing to recognize or internalize that autistic people's conscious experience is separate from their symptoms, which means that c) these allistic people do not recognize that autistic people are people in the same way that they are.

I take this phenomenon to extend beyond autism; I think it is true with regards to the social perception of just about any intellectual or neurodevelopmental disability. The key similarity is that every sort of intellectual/neurodevelopmental disability has its own double empathy problem. Those who are nondisabled, in general, are not able to identify with the conscious experience of what it is like to be intellectually or neurodevelopmentally disabled. Can a person of typical intelligence really understand what it is like to be intellectually disabled? Can a neurotypical person really understand what it is like to be autistic, to have ADHD, OCD, or Tourette's? I doubt it. Meanwhile, can an intellectually disabled person really understand what it is like to have typical intelligence? Can a neurodivergent person really understand what it is like to be neurotypical? I doubt that, too. When people's minds work in fundamentally different ways, it is exceedingly difficult for mutual empathy and identification. It feels impossible to conceive of one's mind working in a way other than it actually does. Otherwise, it feels as though one is not thinking of one's own mind at all. So, while Clare seems right when he claims that people with intellectual disabilities are denied personhood, it is not quite that they are denied personhood because of their lack of intelligence. Rather, their lack of intelligence (or in the case of neurodevelopmental conditions, their different ways of thinking) provides the circumstances that allow them to be denied of their personhood through failures in double empathy.

I want to briefly revisit the discussion of autism to show this sort of non-identification in action. To do so, I will return to Clare's book, *Brilliant Imperfection*, which includes a particularly illustrative example. In one of his arguments about proposed cures to disability, Clare, who is opposed to cure (as will be explored in further detail later), describes how organizations use fear of disability to fundraise for cures. Clare cites two examples of television ads that attempt to do this: one from the Canadian Cystic Fibrosis Foundation, and another from Autism Speaks. The Canadian Cystic Fibrosis Foundation ad capitalizes on the audience's fear of potentially developing symptoms of cystic fibrosis. It uses second person language like, "Cystic fibrosis fills your lungs with fluid, makes every breath a struggle. It's like drowning from the inside."[18] The language asks the audience to imagine what it is like to have cystic fibrosis, and to be terrified of over experiencing those symptoms.

The Autism Speaks ad also uses second person language, but in a very different way. In this ad, a narrator who is meant to serve as a personification of autism, says, "I will rob you and your children of your dreams. I will make sure that every day you wake up, you will cry, 'Who will take care of my child after I die?' And the truth is, I am still winning, and you are scared. And you should be."[19] This ad, called "I am Autism," has become infamous within the autistic community; the Autistic Self Advocacy Network even went as far as to call it "horrifying" for its blatant ableism.[20] But what I want to focus on is who this ad is addressing. While the Canadian Cystic Fibrosis Foundation's ad addresses people who understand that they might one day experience the symptoms of cystic fibrosis, the Autism Speaks ad does not address people who might one day experience the symptoms of autism (as I will discuss later, to do so would be incoherent). Rather, the Autism Speaks ad addresses people who have or might one day have an autistic child or otherwise care for an autistic person, and it asks that audience to be terrified of this eventuality. While the CCFF ad says, in essence, "Wouldn't it be horrible to have cystic fibrosis?" the Speaks ad says, "Wouldn't it be horrible for you to have to deal with an autistic person? Wouldn't it be such a disruption to your life if an autistic person were, unfortunately, to exist?" Both ads are in favor of

[18] Clare, *Brilliant Imperfection*, 89

[19] Clare, *Brilliant Imperfection*, 89

[20] "Horrific Autism Speaks 'I Am Autism' Ad Transcript," Autistic Self Advocacy Network, accessed October 30, 2023 https://autisticadvocacy.org/2009/09/horrific-autism-speaks-i-am-autism-ad-transcript/.

creating a world in which the disability they discuss does not exist, but the ways they address the human element are very different: CCFF advocates for a world in which there is no cystic fibrosis, Speaks advocates for a world in which there are no autistic *people*.

The Speaks ad is troubling on a number of levels, not least of all because it represents an organization that purports to speak for autistic people actively advertising for the eradication of autistic people from the face of the earth. But it, when taken alongside the CCFF ad, is a useful case study. The first thing the Speaks ad shows is how easy, and maybe even natural, it is to treat those with intellectual/neurodevelopmental disabilities as worthless and valueless in the way Clare describes. When organizations like Autism Speaks argue that a world without autistic people in it is a better world than one with autistic people in it, they prove Clare right. But in doing so, they also help reveal a second thing, which is the mechanism by which this process works. The CCFF ad is able to ask its audience to imagine having cystic fibrosis because this is an altogether coherent concept to its presumably nondisabled intended audience. Even though symptoms of cystic fibrosis usually begin in infancy or early childhood, presumably making it difficult for most people with cystic fibrosis to imagine a life without it, it is perfectly coherent to separate a person with cystic fibrosis from their cystic fibrosis. One can conceive of someone with cystic fibrosis suddenly no longer having its symptoms, or someone without cystic fibrosis suddenly developing symptoms.[21] This shows that a nondisabled person is able to imagine themself having cystic fibrosis, meaning that on some level, a nondisabled person can understand that there is something that it is like to be a person who has cystic fibrosis.

The same can be said of other physical disabilities. In some cases, it is not only conceivable, but plausible for someone to imagine developing a physical disability. A person can imagine breaking their spine and becoming paralyzed, or losing the ability to walk due to a condition like multiple sclerosis or Lou Gherig's disease. Indeed, most of us probably know someone who has experienced what it is like to become disabled, and many of us have experienced it firsthand. If it is so easy for a nondisabled person to imagine that they could be a physically disabled person, even if the experience of physical disability that they imagine is wholly inaccurate, then they must understand that there is something that it is like to be a physically disabled person. The point here is that while there seems to be a double-empathy problem preventing nondisabled people to understand that there is something that it is like to be an intellectually or neurodevelopmentally disabled person, such an epistemic problem does not seem to exist for physical disabilities, or at least does not exist to the same extent.

This facet of disability has an important practical implication with regards to our politics of cure. While becoming disabled certainly represents a significant change in one's identity, it does not represent the cessation of identity, or the replacement of identity with another identity, as evidenced by the fact that nondisabled people can imagine that *they themselves* could be disabled. A person who becomes paralyzed, though they have changed in a significant way, is still the same person. Because of this, it is also perfectly conceivable to envision a cure for a physical disability that does not destroy identity. If identity can be preserved between physical embodiments, then surely, it can survive the loss of disability just as it can survive the acquisition of disability. I do not think this is true of intellectual/neurodevelopmental disabilities like autism, though, as evidenced by allistic and otherwise nondisabled people being unable to imagine that *they themselves* could be intellectually or neurodevelopmentally disabled. For example, although the conscious experience of being autistic is separable from autism symptoms (even though the symptoms are results of the conscious experience), that experience of being autistic is not separable from identity. We can imagine someone who displays many of the traits most associated with autism (society isolation, hyperfixations, sensory overstimulation, etc.), but who is not autistic. These traits could be the result of other neurodevelopmental conditions, they could be psychologically acquired as a result of something like trauma, or maybe this person just happens to have this sort of personality. Similarly, we can imagine an autistic person who, though years of practice and possibly through applied

[21] This is perhaps an implausible scenario, but it is a conceivable, and therefore, a coherent one.

behavior analysis (ABA) therapy,[22] has learned to mask their symptoms to the point that they are no longer recognizably autistic. However, such a person has not ceased to be autistic. While physical disability is ultimately defined by symptoms, which are deeply connected to identity, but ultimately separable from it, intellectual/neurodevelopmental disability is defined not by symptoms, but by identity itself.

This, I think, is the most salient difference between physical disabilities and intellectual or neurodevelopmental disabilities, and I consider this difference essential to how we should approach the politics of cure. I want to partially react against strong anti-cure narratives like those presented by Eli Clare with regards to physical disabilities, although there is much in these narratives that I find compelling. I agree with Clare that cure is not essential for those who live with physical disabilities. It seems undeniable that many physically disabled people find identity, solace, and joy in their disabilities, and that living with a physical disability does not mean that one must have a lessened quality of life. I also agree with Clare cure should not be the priority for the disability rights movement. I think Clare is quite correct that what disabled people need most is not cure, but civil rights. So, I agree with Clare that it is perfectly legitimate, and even admirable, for physically disabled people not to desire a cure for their conditions, and to fight against resources being allocated for the creation of one at the expense of civil rights. However, what I want to argue against is the notion that when a physically disabled person desires a cure, that it represents some sort of internalized ableism that is the product of social injustices. Just as I think it is perfectly legitimate for a physically disabled person to reject a cure, I think it is perfectly legitimate for a physically disabled person to desire one. In principle, I do not think there is anything wrong with a world in which cures for physical disabilities exist, there is only something wrong with a world in which they are required.

However, I do think there is something wrong with a world in which cures for intellectual/neurodevelopmental disabilities exist because such a word makes it possible to eradicate the identities of people with such disabilities. When we propose that we cure someone of a condition like intellectual disability or autism, we are proposing something incoherent because that person's identity cannot survive cure in the way that a physically disabled person's identity can. Curing someone's intellectual/neurodevelopmental disability is akin to destroying that person's identity and replacing it with a new one. One may just as well kill an intellectually/neurodevelopmentally disabled person and replace them with a new nondisabled person. As such, I do think that when an intellectually/neurodevelopmentally disabled person desires a cure, as many do, it is the result of internalized ableism. It is an instance of internalizing the social perception of oneself as lacking value or worth. I see it as a type of suicidality; it is the idea that one might as well destroy oneself because one was never a person to begin with. As such, we should in our politics reject cure for intellectual and neurodevelopmental disability. We must not allow these identities to be eradicated.

I bring up the politics of cure for two reasons. Primarily, it seems to me that the position on cure that arises from my analysis of social ascriptions of personhood is of intrinsic interest. But, secondarily, and perhaps more relevantly, I think it shows how integral an account of personhood can be to this position. As such, my response to Clare's politics of cure represents more than a minor variation. Though I mostly agree with Clare's view on the generals of social personhood, and I completely agree with his view of who we must include, the differences in the mechanisms of personhood that I illustrate in my account result in a view of cure that deviates substantially from Clare's. So, I want to conclude on the note that the discussion of personhood cannot be confined to the margins of our discourse on disability. Personhood, with its complexities, is an issue of central importance.

[22] For the record, I do not support ABA; autistic people subjected to ABA have described it as "compliance training," and a "Pavlovian torture method that attempts/succeeds in removing an autistic's brain functions and replaces their normal functions with those of the dominant culture." See Therese M Cumming et al., " 'I Was Taught That My Being Was Inherently Wrong': Is Applied Behavioural Analysis a Socially Valid Practice?," *International Journal of Arts, Humanities, and Social Sciences Studies* 5, no. 12 (December 2020).

## Bibliography

Clare, Eli. *Brilliant Imperfection: Grappling with Cure*. Durham: Duke University Press, 2017.

Cumming, Therese M, Iva Strnadová, Joanne Danker, and Caroline Basckin. " 'I Was Taught That My Being Was Inherently Wrong': Is Applied Behavioural Analysis A Socially Valid Practice?" *International Journal of Arts Humanities and Social Sciences Studies* 5, no. 12 (December 2020): 72–82.

Gruen, Lori. "The Moral Status of Animals." *Stanford Encyclopedia of Philosophy*, June 23, 2021. https://plato.stanford.edu/entries/moral-animal/.

"Horrific Autism Speaks 'I Am Autism' Ad Transcript." *Autistic Self Advocacy Network*, September 23, 2009. Accessed October 30, 2023. https://autisticadvocacy.org/2009/09/horrific-autism-speaks-i-am-autism-ad-transcript/.

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

Koterski, Joseph W. "Boethius and the Theological Origins of the Concept of Person." *American Catholic Philosophical Quarterly* 78, no. 2 (2004): 203–24. https://doi.org/10.5840/acpq200478212.

Milton, Damian E. M.. "On the ontological status of autism: the 'double empathy problem.' " *Disability & Society* 27, no. 6 (2012) 883-887. Accessed October 30, 2023. https://www.tandfonline.com/doi/full/10.1080/09687599.2012.710008.

Nagel, Thomas. "What Is It Like to Be a Bat?" *The Philosophical Review* 83, no. 4 (October 1974): 435. https://doi.org/10.2307/2183914.

Taylor, Charles. *The Concept of a Person*. Cambridge, UK: Cambridge University Press, 1985.

"What Are the Symptoms of Autism?" *Autism Speaks*. Accessed October 30, 2023. https://www.autismspeaks.org/what-are-symptoms-autism.

# 'Fearish' Moods: A Non-Intentional Theory

Krisztian Kos, *University of St Andrews*

Are moods directed at objects? Many philosophers have answered 'yes': moods are about things like events or people – their intentional objects. One intentionalist view of moods – which takes them as directed at their objects – is put forward by Carolyn Price, in her "Affect Without Object: Moods and Objectless Emotions", where she claims that an apprehensive mood is about how likely it is that a threat will occur. In this essay, I will develop some of Price's insights and use them to give a non-intentionalist account of moods. In Section 1, I first characterise moods and contrast them with emotions. Next, in Section 2, I outline Price's intentionalist theory of moods and raise two problems with it. Her theory of moods does not sufficiently account for how they function and mischaracterises their motivational aspect. Then, in Section 3, I propose a new way of thinking about moods by drawing on a theory of colour-blindness. Byrne and Hilbert's 'alien view' treats the colours that colour-blind people see as less determinate and fine-grained than the ones that people with regular vision see. After drawing the relevant parallels in the case of moods and emotions, I show how moods, on this account, are pre-intentional, rather than non-intentional mental states and I finish, in Section 4, by addressing some objections to this view. I conclude that the mood of a subject should not be thought of as belonging to the same class as (intentional) emotions, since moods are pre-intentional states that structure the space of possible mental states in virtue of determining how likely it is that we experience some intentional state.

## 1 Characterising moods

Moods make up an important part of our affective experiences – those which are related to feelings and emotions. Moods are usually seen as prevailing attitudes that determine the way we generally feel, usually brought about by our environment. Paradigmatic examples of moods are depression and happiness which have widespread effects on our beliefs, desires and actions. In a depressed mood, I perceive everything in my surrounding environment as bleak and devoid of significance and form negative emotions, preventing me from pursuing anything. An elated mood, however, presents my surroundings to me as open and filled with new opportunities, making me feel more confident and open to trying out what my environment offers me.

Moods have pervasive effects on our inner life. They spread to all parts of our experience and permeate our experience of the world. An acute mood like anxiety can overwhelm us, profoundly impacting the way we attend to our current situation, while 'we just can't shake'[1] a mood like melancholy, which is a pensive state characterised by feelings of sorrow and 'less intense than grief'.[2] Moods come in varying intensities and have a global feature to them,[3] in that they affect our thought-processes, our emotions, and our behaviour.

Finally, moods have been seen as not directed towards anything specific. So, moods are usually thought of as intentional states – states which are about or directed at something. And the

[1] Laura Sizer, "Towards a Computational Theory of Mood," *British Journal for the Philosophy of Science* 51, no. 4 (December 2000): 743. https://www.jstor.org/stable/3541726

[2] Mathea Slåttholm Sagdahl, "Melancholy as Responding to Reasons," *International Journal of Philosophical Studies* 29, no. 3 (July 2021): 334. DOI: 10.1080/09672559.2021.1936120

[3] See Jonathan Mitchell, "The intentionality and intelligibility of moods," *European Journal of Philosophy* 27, no.1 (July 2018). https://doi-org.ezproxy.st-andrews.ac.uk/10.1111/ejop.12385

intentional objects of moods – that which they are about – tend to be more general, as opposed to the more specific objects of emotions. Whereas we tend to be angry *at* some person, or joyful *about* an upcoming event, we do not say that we are in a depressed mood about anything in particular, but rather about our general situation. Some have claimed that moods have general intentional objects like one's situation,[4] and others have said that moods are directed at the one's total environment.[5] This contrasts with the more specific and local objects of emotions, like particular events. Moreover, Carolyn Price[6] has claimed that whereas emotions are about specific occurrences having taken place, moods are concerned with the likelihood that some event will happen.

## 2 Price's Intentionalist View of Moods

On Price's intentionalist account – where moods have intentional objects they are about – moods are states of vigilance that look out for things that might trigger some mood. They are intentional states, in that they are directed at 'how things are likely to turn out'[7] in one's situation. More specifically, moods are signals that carry information as to how likely an event that causes a mood will happen. The function of an apprehensive mood, then, is to 'adapt the subject to an environment in which there is an increased probability that… a threat will occur'.[8] On this view, the intentional object of a mood X concerns the likelihood that a situation that triggers X will happen.

Moods function to adapt a subject to an environment where a relevant trigger of a mood is more likely to happen. Price also points out that this 'need not imply that these moods typically succeed in performing this function'.[9] Rather, moods need only perform this function sometimes, just enough to improve a subject's well-being.

Next, Price draws an important distinction between the descriptive and directive contents of mental states. Emotions have both descriptive content – the information they usually carry – and directive content – a result to be attained. They signal that an event has occurred and motivate 'actions designed to cope with [a] situation'.[10] Moods, however, possess only descriptive content, since 'it is not the function of these signals to motivate the subject to act'.[11] All that an apprehensive mood does, for instance, is make the apprehensive subject 'poised to flee'[12], but without creating an urge to flee. So, moods, which lack directive content, do not motivate actions. Once again, Price makes this claim 'without supposing that moods always or typically carry this information'.[13]

However, I argue that there should be some relationship between what a mood brings about and its success in performing its function. Let us consider an apprehensive mood, the function of which is to adapt the subject to an environment where there is a higher chance that a threat will happen. In a fearful mood, we tend to narrow our focus on to the most dangerous threat and pay less attention to smaller details. The function of fear – to adapt us to an environment where some danger is more likely to occur – is not successfully realised, for a small range of concentration and a lack of attention makes us more susceptible to the other potential dangers we have not considered or are biased against.

Furthermore, even if we grant that moods do not need to typically satisfy their function, they should at least *aim* to perform their function successfully. Experiencing an apprehensive mood

[4] See Peter Goldie, *The Emotions: A Philosophical Exploration* (Oxford: Oxford Academic, 2002). https://academic.oup.com/book/27031 and Robert Solomon, *The Passions: Emotions and the Meaning of Life* (Indianapolis: Hackett Publishing, 1993).

[5] See Mitchell, "Intentionality and intelligibility".

[6] See Carolyn Price, "Affect Without Object: Moods and Objectless Emotions," *European Journal of Analytic Philosophy* 2, no.1 (November 2006). https://hrcak.srce.hr/file/135314

[7] Price, "Affect Without Object", 65.

[8] Price, "Affect Without Object", 57.

[9] Price, "Affect Without Object", 57.

[10] Price, "Affect Without Object", 54.

[11] Price, "Affect Without Object", 61.

[12] Price, "Affect Without Object", 63.

[13] Price, "Affect Without Object", 59.

should aim to adapt us to a situation where a threat is more likely to occur, but we often try to suppress our fear and attempt to calm ourselves down. Even when we try to calm down because we are having conflicting moods or experiencing mild apprehension, we are trying to restore a calm mood – which allows us to process information more carefully – and avoid an agitating mood, like apprehension. By returning to our usual, day-to-day mood, we attempt to straighten out our reasoning and find the best way out of this threatening situation. An apprehensive mood – by itself or in conflict with others – that excites us and makes us 'overreact' does not seem to help us in dealing with our threatening situation and its effects on us.

Next, let us suppose that there is no significant relationship between the information that a mood carries and the kind of mood it is. A specific mood, like irritability, then, is not linked to the information that it carries – that an offence is likely to occur. Since irritability need not typically carry such information, and does not do so, we can infer that the content of this mood does not play much role in determining the nature of the resulting irritable mood. However, if this is the case, then how does a situation warrant one mood rather than another? Consider a situation where we think it is likely that the person we are talking with aims to provoke us in some way. We believe that the chances that we will be offended is high. So, we are put in some mood Y whose descriptive content carries the information that an offence is more probable to occur. In this case, then, an irritable mood matches mood Y more – or, is more appropriate – than, for example, a joyful mood. This is because the evaluative aspect of the descriptive content – that which carries a positive or negative value and makes the mood feel good or bad – favours a mood with a corresponding evaluative aspect. In this case, the descriptive content has a negative evaluative dimension which matches the negative experience of an irritable mood more than the positive experience of a joyful mood. If we repeatedly find ourselves in a situation with such a trigger present, then the descriptive content of the mood we are in will typically have associated with it an irritable mood. So, there seems to be some significant relationship that holds between the descriptive content of a mood and its nature.

Now, let us suppose that there *is* a significant relationship between the information that a mood carries and the nature of the mood. Here, moods do typically carry the information that a relevant trigger is likely to occur. One way they can do so is by representing the world in a reliable way across a range of contexts. So, if an irritable mood typically carries the information that an offence is likely to occur, then it represents situations where an offence did occur in an accurate way – e.g. in all situations where we were in an irritable mood, there were more cases where an offence did happen than cases where no offence took place. But, if an irritable mood typically carries the relevant information, then, upon encountering a new situation where there is a high chance of an offence, it is hard to see how we will not be motivated to act in some way – for example, to adapt to our environment. If we have performed certain actions in the past based on the correct information carried by our irritable mood in many cases – perhaps leading to successful outcomes like dealing with our situation effectively – then we will be more inclined to perform those actions in a new situation where we are presented with similar information via our irritable mood. Therefore, there being a significant relationship between the kind of mood we are in and the information it carries means that moods can be motivating.

We can also see this motivating feature of moods when Price considers the difference between moods and objectless emotions – emotions that are not about anything or seem to be directed at nothing, like happiness immediately after waking up. She considers the example of apprehension and objectless fear. When we are walking on a deserted street at night and a sudden feeling of fear comes over us – without knowing what it is directed at – the main difference between the two states, she claims, is that objectless fear can motivate action, while apprehension cannot. While a subject experiencing an objectless fear can have an 'urge to flee, or… be torn between fleeing and hiding', an apprehensive subject would merely be 'poised to flee'.[14] When the subject has an objectless fear, they are still 'motivated to act in some way consistent with [their] fear',[15] despite

[14] Price, "Affect Without Object", 63.
[15] Price, "Affect Without Object", 63.

there not being enough information to decide how to act ('torn' between alternatives). For an apprehensive subject, though, there are no alternatives presented *at all*. They are merely 'poised to flee', and not torn between opposing motivations (since there are none).

But, just because there is no determined action that they are drawn to, like fleeing, does not mean that the subject is not motivated to take themselves out of the threatening situation. Imagine, for instance, someone who is having an unproductive day who suddenly decides to stop wasting it and do something worthwhile their time. To motivate action, they do not need to know what their options are or how they want to act. It is enough, as a motive for them, to want to escape the situation without having something to move toward. Similarly, an apprehensive mood can motivate us to escape the situation, without determining any of our possible routes of escape (and without presenting alternatives for *how* to escape). On this motivational understanding of moods, we can construct 'poised to flee' to mean that we want to leave the situation at hand, but without knowing how to do so, rather than the situation merely causing us to be 'poised to flee'.

## 3 Colour-blindness and moods

Instead of thinking of moods as intentional mental states, I propose an alternative way of thinking of moods. First, I will present a model of colour-blindness that can serve as a tool for thinking about moods and their relation to emotions. Then, building on some of Price's insights and Matthew Ratcliffe's theory of existential feelings, I outline how to apply this model to moods.

### 3.1 Byrne and Hilbert's theory of colour-blindess

Regular human vision consists of short wave, medium wave, and long wave cone cells. These cone cells combine to output colors along three corresponding opponent processes: black-white, red-green, and blue-yellow. These make up our colour space – all the colours that we can see (see Figure 1). The 'response to the colour at one end of the dimension is antagonistic to the response at the other end',[16] meaning that an increase in the detection of red means a decrease in the detection of green, and detecting unique yellow – yellow with no traces of any other colour – entails zero detection of unique blue.

On Byrne and Hilbert's view of colour-blindness,[17] colour-blind people detect different colours from the ones that people with regular vision detect. So, dichromats – colour-blind people who are missing one of the opponent processes, like the red-green one – do not see only a subset of the colors that trichromats see (like the space of only lighter and darker shades of yellow and blue). Rather, they 'see some other colours entirely – some less determinate colours' [18] that are less specific and less fine-grained than the colours that trichromats see.[19]

According to Byrne and Hilbert, when a person with regular color vision (a trichromat) sees unique yellow, for example, then the red-green opponent process gives a neutral signal, i.e. that there is a balance between red and green. Dichromats, however, 'have no functioning red-green channel',[20] so there is nothing to signal that there is a balance between red and green. These two cases are different, though, because a neutral signal is not the same as no signal: the former accurately indicates unique yellow, whereas the latter produces no signal to accurately determine unique yellow. So, what is signaled for a red-green colour-blind person when what we would normally call yellow is present? The output of their yellow-blue channel signals '*yellowishness*': a color that is somewhat yellow and lies between unique red and green on the yellow half of the hue

[16] Fiona Macpherson, "Novel Colour Experiences and Their Implications," in The Routledge Handbook of Philosophy of Colour, eds. Derek H. Brown and Fiona Macpherson (London: Routledge, 2020), 180. https://doi-org.ezproxy.st-andrews.ac.uk/10.4324/9781351048521

[17] See Alex Byrne and David R. Hilbert, "How Do Things Look to the Color-Blind?" in Color Ontology and Color Science eds. Jonathan Cohen and Mohan Matthen (Cambdridge, MA: MIT Press Scolarship Online, 2013). https://doi.org/10.7551/mitpress/9780262013857.003.0012

[18] Macpherson, "Novel Colour Experiences", 184.

[19] Macpherson, "Novel Colour Experiences", 184.

[20] Byrne and Hilbert, "How Do Things Look to the Color-Blind?", 282.

circle. Such dichromats cannot see determinate colours like unique yellow and orange. Instead, they 'see colours that are less determinate than people who are not colour-blind'[21] – colours like *yellowish* and *bluish*.

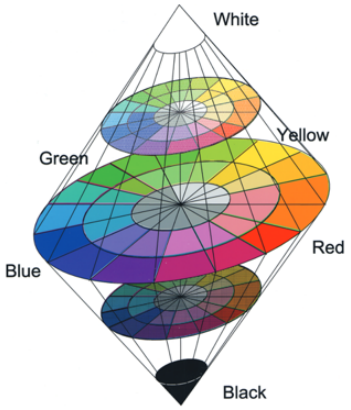


Figure 1: The Classical Colour Space, with red-green and blue-yellow opponent processes along the surface of the circles. Taken from Churchland (2005), p.531 [23].

### 3.2 '*Fearish*' moods

Byrne and Hilbert's model of colour-blindness can help us understand moods and emotions. Price suggests a way to think about the relationship between the two: a mood increases a subject's sensitivity towards certain kinds of cues or triggers that help generate a relevant emotion.[24] Suppose, for example, that a subject is in an apprehensive mood. So, when this fearful subject hears the wind blowing violently against their window, they interpret this event as having a higher chance of threatening them. Hence, the subject needs less 'in the way of additional evidence to warrant'[25] their fright about the wind outside. The mood of the subject, then, makes them more sensitive towards relevant emotions.

Let us now think of emotions like the determinate colours that trichromats see, and of moods like the less determinate colours that dichromats see. Emotions are, then, more fine-grained than moods since they are specific mental states. This seems consistent with the pervasive quality of moods – while emotions tend to be about single objects, moods spread to all parts of our experience and, hence, allow a range of emotions to form. Conversely, moods are more general and range over different contexts, being less precise than emotions. Moods have often been described as hazy,[26] having 'vague, nebulous characters'[27] and changing the experience of our environment 'in ways that are difficult to pin down'.[28] In some mood, it is also more likely that we develop a range of emotions. In a joyful mood, we can be excited about an event in the future happening, satisfied with our current situation, and elated about good news we have received.

To continue, consider the diagram in *Figure 2*.[29] This diagram shows emotions spread out around a circle divided up along two axes: the horizontal ranging from pleasant to unpleasant, and the vertical ranging from active to deactivated. Emotions are grouped according to how they rank

[21] Macpherson, "Novel Colour Experiences", 185.
[24] Price, "Affect Without Object", 64.
[25] Price, "Affect Without Object", 64.
[26] Goldie, "The Emotions".
[27] Sizer, "Computational Theory", 765.
[28] Matthew Ratcliffe, "The Phenomenology of Existential Feeling," in Feelings of Being Alive eds. Joerg Fingerhut and Sabine Marienberg (Berlin, Boston: De Gruyter, 2012), 34. https://doi.org/10.1515/9783110246599.23
[29] The following summarises ideas based on James A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology* 39, no. 6 (1980). DOI: 10.1037/h0077714

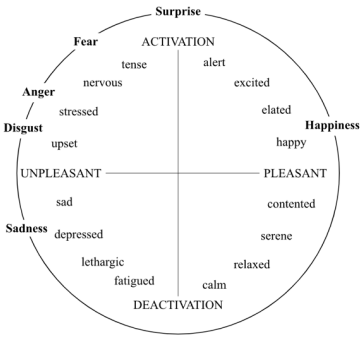


Figure 2: Circumplex structure of emotions. Taken from Scarantino and de Sousa (2021).[30]

on each axis. Elation is a very positive emotion where we tend to be more energetic than average, while depression is characterised as a negative emotion (which, I will argue, should be revised) where we tend to be more inactive.

Using the way that Byrne and Hilbert view colour-blindness, and our previous comments on moods and emotions, we can think of moods as grouping together relevant emotions. Whereas the domain of emotions is more detailed, moods are less determinate and less fine-grained than emotions. To illustrate, imagine a more accurate diagram. Considering the part of the circle where we see 'nervous' and 'tense', we would see a host of other emotions there as well – emotions like 'worried', 'anxious', 'alarmed', 'frightened', 'panicked' and 'horrified'. These emotions would spread out across this area according to how unpleasant and active they are. 'Panic' would be closer to where 'tense' is, whereas 'worried' would be closer to 'nervous'. All these emotions, as the diagram shows, are grouped under 'fear'. When we claim that we are 'anxious', we can determine our exact location on this diagram: the position of 'anxious' according to how it ranks along the two axes. If, however, we say that we are in an apprehensive mood – that we are fearful – our mental state is less determinate, less specific and more vague. We are not at an exact location on this diagram, but rather our mental state concerns the range of emotions that fall under 'fear'. My suggestion is that this mental state that we denote by 'fear' is best understood as a fearful, or following Byrne and Hilbert's terminology, '*fearish*' mood.

Furthermore, this way of thinking of moods can make this diagram more accurate. Depression is usually considered a mood characterised by a pervasive and 'indiscriminate generality'[31] with the 'typical responses found in depression [including] feeling…miserable, dispirted, listless'.[32] So, by treating depression as a less determinate mood located below sadness and connected to several emotions, like feeling fatigued and listlessness, we can make more sense of its nature than seeing it as one specific emotion.

This view is also consistent with Price's suggestion that moods make us more sensitive towards certain kinds of cue. In a '*fearish*' mood, we are more sensitive towards the occurrence of a threat and less likely to miss one. Once such a threat occurs, we develop an emotion towards it, like worry, covered by this mood. So, in a '*fearish*' mood, we are more likely to experience emotions like anxiety, and less likely to experience emotions like elation (which would fall under a '*happy-ish*' mood).

So, undergoing a mood is a less determinate experience where experiencing some relevant emotions is more likely. The experience of a '*fearish*' mood is a more indeterminate and vague experience that ranges over a host of emotions – like worry and horror – situated between the unpleasant and active poles in Figure 2, analogous to a dichromat's less determinate experience of '*yellowish* that lies between unique red and green. By increasing the likelihood of experiencing such

[31] Jennifer Radden, "The Self and Its Moods in Depression and Mania," *Journal of Consciousness Studies* 20, no. 7-8 (January 2013): 84.

[32] Radden, "The Self", 84.

emotions – by making us more sensitive to certain cues, for instance – this mood involves a broader experience that concerns this set of related emotions – like how '*yellowish*' would be related to unique yellow and orange. In a mood, we are undergoing an experience where we are more likely to experience certain emotions rather than others. In a '*fearish*' mood, like when walking on a deserted street at night, we are more likely to feel worried about the figure approaching us and horrified at the dark side streets. In turn, this experience of a '*fearish*' mood is a less determinate one that ranges over (but is not identical to any of) the related emotions.

### 3.3 **Pre-intentionality**

Drawing on Matthew Ratcliffe's work on existential feelings can help to see more clearly why moods are not intentional states. In this section, I show how moods are 'pre-intentional' – rather than intentional – states that determine the kinds of emotions we can experience.

For Ratcliffe, everything we experience is 'permeated with fundamental affectivity',[33] meaning that basic affective states – states laden with feeling – 'pre-structure our experience'.[34] These basic states – existential feelings – are 'background orientations' that influence the 'ways in which specific objects can appear to us'[35] and the kind of intentional states we can experience. They 'shape our space of possibility'[36] by determining what experiences are possible for us. For example, a feeling of detachment structures our space of possible experiences by making us feel hopeless about our situation, while making it harder to feel excited about any events. 'Intentional states presuppose existential feelings' because to feel threatened by an event, for example, 'one's world must accommodate possibilities of those kinds'.[37] Existential feelings are 'pre-intentional',[38] since they precede intentional states in virtue of determining 'what kinds of intentional state are amongst one's possibilities'.[39] An existential feeling is not about or directed at anything, but is an 'underlying tone'[40] that determines how and what kind of intentional states we experience. The existential feeling of a depressed person, for instance, makes salient the emotion of feeling helpless about oneself and the belief that one's actions are worthless.

On the view suggested, moods are similar to Ratcliffe's existential feelings. In a '*fearish*' mood, we are more vigilant as to the occurrence of a threat and more likely not to miss any signs of danger. From Section 3.2, this entails that we are more likely to experience emotions like anxiousness or fright, since we are on the lookout for signs that give rise to such emotions. This mood, then, structures the space of emotions that we can have. So, the kinds of emotions we will and will not have in a '*fearish*' mood will be different from the ones in a '*happy-ish*' mood – where there is a higher chance of experiencing elation or joy, and a lower chance of experiencing anxiousness or fright. Depending on the mood we are in, the space of emotions that we can experience changes. Therefore, moods are not themselves directed at anything, but are 'pre-intentional' in virtue of determining what kinds of emotions are in our space of possibility.

The fact that one can experience a pre-intentional mood while simultaneously having an emotion, without these two states being identical, helps to see the distinction between the two more clearly. Consider someone who is expecting some good news that is likely to arrive. There is a high chance that a positive trigger will happen, and so the subject is in an excited mood. Simultaneously, the subject can feel content with the view from their window, since their excited mood, in virtue of increasing the likelihood of undergoing contentment, can bring about this emotion. However, there is still a difference between the pre-intentional mood that makes it more likely to experience certain emotions (like contentment) and constitutes a broader experience, and the single localised and intentional emotion of contentment.

[33] Kreuch, "Existential Feelings", 75.
[34] Kreuch, "Existential Feelings", 85.
[35] Kreuch, "Existential Feelings", 82.
[36] Kreuch, "Existential Feelings", 82.
[37] Ratcliffe, "The Phenomenology of Existential Feelings", 32.
[38] In both Kreuch, "Existential Feelings" and Ratcliffe, "The Phenomenology of Existential Feelings".
[39] Ratcliffe, "The Phenomenology of Existential Feelings", 32.
[40] Kreuch, "Existential Feelings", 82.

Thinking of moods as existential feelings also makes more salient the way they both structure and constitute experience, as mentioned in Section 3.2. Moods are 'a pre-structuring background of all experience' – in virtue of determining which emotions we are more (and less) likely to experience – and 'a part of experience at the same time'[41] – in virtue of being a less determinate experience that ranges over the related set of emotions.

## 4 Objections

In this section, I consider some objections to the view proposed in Section 3.

One might object that thinking of moods as determining what emotions we can and cannot have is an inaccurate picture of how moods work. Consider a subject in a miserable mood. If this mood structures their space of possible emotions, then emotions like melancholy and agony will be in the range of possible emotions, whereas cheerfulness and bliss will be excluded. But, when someone attempts to cheer up this subject – by consoling them and trying to make them laugh – then if cheerfulness is not in the range of possible emotions the miserable subject can experience, then this means that they cannot be cheered up: they will only be capable of experiencing '*miserable-ish*' emotions. However, after a while, initially miserable subjects can start to cheer up and have their mood improve, so miserable subjects *can* experience positive emotions like cheerfulness. Therefore, the objection concludes, thinking of moods as determining what emotions we can and cannot possibly have is not consistent with how moods really work.

However, following from the discussion above, we can think of moods not as affecting the *possibility* of experiencing some emotion, but rather as affecting the *likelihood* of experiencing some emotion. Thinking of moods in this way still shows that moods, since they are pre-intentional, structure the space of possible emotions. But, instead of determining which emotions are possible (and not), moods determine which emotions are more (and less) likely to be experienced by a subject. To continue, they are still pre-intentional states that are not directed at anything, but rather influence the chances of undergoing some emotion(s) rather than others. That in a miserable mood we can experience cheerfulness shows that the miserable mood is distinct from an intentional emotion of misery where we would experience only that one, specific emotion. On the view that moods determine likelihood instead of possibility, the theory proposed is consistent with how moods work in reality, while preserving the idea that moods are pre-intentional states that structure the space of possible emotions – by making the experience of some more likely than others.

One might also object that since moods allow opposing emotions to occur, they do not succeed in performing their function – namely, to adapt the subject to an environment where the relevant trigger has a high chance of occurring – and that this account falls prey to the same objection presented against Price's account in Section 2. However, on the account I have proposed, moods do *typically* succeed in performing their function. By increasing the sensitivity towards certain kinds of cue, and thereby making it more probable that relevant emotions will obtain, moods give rise to the relevant emotions more often than they give rise to opposing emotions – typically satisfying their function.

A final objection may point out that since moods do not have directive content and do not seem to have descriptive content on this pre-intentionalist view, moods cannot motivate actions. Arguing that moods do or do not have descriptive content is outside the scope of this paper, but showing that moods can be motivating does not hinge on this matter. This is because merely in virtue of increasing the likelihood of experiencing some emotions, a given mood will have associated with it relevant emotions with directive content. So, by giving rise to directly motivating emotions, moods incline us towards performing certain actions which make moods (indirectly) motivating.

[41] Kreuch, "Existential Feelings", 85.

## 5 Conclusion

In this paper, I have presented a non-intentionalist account of moods. First, I characterised moods, highlighting their pervasiveness. Then, I outlined Price's account of moods as states of vigilance that concern the likelihood of relevant triggers occurring in a subject's environment. I presented some problems with this view, and then used some of Price's insights to put forward my own account of moods. Drawing on a theory of colour-blindness, I showed that experiences of moods are less determinate than the more specific emotions that – in virtue of making it more likely that we experience them – they group together. In a '*fearish*' mood, we have a more vague experience where we are more likely to experience emotions like worry and anxiety. I then went on to show how these moods are pre-intentional states – structuring the range of emotions we experience – and finished by addressing some objections. I conclude that such a non-intentional account of moods is a useful way to think about them, and should be explored further.

## Bibliography

Byrne, Alex and David R. Hilbert. "How Do Things Look to the Color-Blind?" *Color Ontology and Color Science*, ed. Jonathan Cohen and Mohan Matthen, 259-290. Cambdridge, MA: MIT Press Scolarship Online, 2013. https://doi.org/10.7551/mitpress/9780262013857.003.0012.

Goldie, Peter. *The Emotions: A Philosophical Exploration*. Oxford: Oxford Academic, 2002. https://academic.oup.com/book/27031.

Kreuch, Gerhard. "Matthew Ratcliffe's Theory of Existential Feelings." In *Self-Feeling*, 73-99. Cham: Springer Cham, 2019. https://doi.org/10.1007/978-3-030-30789-9.

Macpherson, Fiona. "Novel Colour Experiences and Their Implications." In *The Routledge Handbook of Philosophy of Colour*, ed. Derek H. Brown and Fiona Macpherson, 175-209. London: Routledge, 2020. https://doi-org.ezproxy.st-andrews.ac.uk/10.4324/9781351048521.

Mitchell, Jonathan. "The intentionality and intelligibility of moods." *European Journal of Philosophy* 27, no.1 (July 2018): 118-135. https://doi-org.ezproxy.st-andrews.ac.uk/10.1111/ejop.12385.

Price, Carolyn. "Affect Without Object: Moods and Objectless Emotions." *European Journal of Analytic Philosophy* 2, no.1 (November 2006): 49-68. https://hrcak.srce.hr/file/135314.

Radden, Jennifer. "The Self and Its Moods in Depression and Mania." *Journal of Consciousness Studies* 20, no.7-8 (January 2013): 80-102.

Ratcliffe, Matthew. "The Phenomenology of Existential Feeling." In *Feelings of Being Alive*, ed. Joerg Fingerhut and Sabine Marienberg, 23-54. Berlin, Boston: De Gruyter, 2012. https://doi.org/10.1515/9783110246599.23.

Russell, James A. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* 39, no. 6 (1980): 1161-1178. https://doi.org/10.1037/h0077714.

Sagdahl, Mathea Slåttholm. "Melancholy as Responding to Reasons." *International Journal of Philosophical Studies* 29, no. 3 (July 2021): 331-350. https://doi.org/10.1080/09672559.2021.1936120.

Scarantino, Andrea and Rondald de Sousa, "Emotion," in *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), ed. Edward N. Zalta. Accessed December 31, 2023. https://plato.stanford.edu/archives/sum2021/entries/emotion/

Sizer, Laura. "Towards a Computational Theory of Mood." *British Journal for the Philosophy of Science* 51, no.4 (December 2000): 743-769. https://www.jstor.org/stable/3541726

Solomon, Robert. *The Passions: Emotions and the Meaning of Life*. Indianapolis: Hackett Publishing, 1993.

# The Moral Character of Mental Illness

Rohan Mavinkurve, *University of St Andrews*

This paper offers a reimagination of Thomas Szasz's claim that mental illness is a myth. His idea that mental illness actually constitutes moral problems is expanded upon with a novel moral framework that makes the claim easier to grasp and advocate for. The argumentative strategy used is intended to bypass the major extant debate about the scientific validity or natural kind status of mental illnesses. Szasz's selective elimination of mental but not physical illnesses is vindicated via an epistemic reduction of the mental features to moral features, which does not parallelly obtain for physical features. This solution is optimised to address the criticisms of R.E. Kendell, arguably Szasz's foremost critic.

## 1 Introduction

The rejection of the psychiatric category of mental illness is often sloganised as 'mental illness is a myth'. This is credited to Thomas Szasz, who claimed that mental illness is not a legitimate category of illness in the way that physical illness is. He argued that 'mental illness' is a metaphor for moral problems, which have been mistaken for medical problems[1].

The aim of this paper is to reimagine Szasz's goal with an expansion on the morality claim that clarifies major criticisms of Szasz. I argue against two consensus ideas of mainstream psychiatry: first, that mental illnesses form a legitimate category of illness, and second, that they are distinct from mental responses that are expected and culturally sanctioned responses to external factors, in that they must arise from a dysfunction in the individual. The second idea is the DSM definition of mental illness, its most influential understanding[2]. This is the definition of mental illness I will attribute to the opposing view, and seek to refute. If I am successful in dispelling it, the rejection of the first idea should follow.

I argue that the concept of mental illness tracks something belonging to the moral realm. This realm is populated with other conditions not of clinical interest; 'mental illness' is not unique or meaningful as a category. The claim, then, will turn out to be true if there are insufficient grounds (metaphysically or epistemically, as will be seen) for identifying mental illness as a subset of the set of moral problems. I do not commit to all tenets of Szasz's thought – the goal is only to uphold the non-status of mental illness.

The plan for the paper is as follows. Section 2 reviews core ideas of Szasz with associated remarks (not exhaustive of either Szasz or his critics, but sufficient for this paper). Section 3 evaluates Hanna Pickard's defence of Szasz, which is of partial interest to this one. Section 4 suggests a moral framework informed by Aristotelian ethics that better captures Szasz's argument. Sections 5 provides a proposal that recasts illness on a moral dimension, denying that it is a meaningful metaphysical category. Section 6 then elucidates how the discriminatory elimination of just mental illness can be achieved without relying on metaphysics. This is done in addressing R.E. Kendell's challenge to Szasz, with a discussion of Szasz's response and how it is vindicated on the presently suggested conception of morality. Section 7 considers objections to the argument, especially clarifying the argument in section 6.

[1] Thomas S. Szasz, *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*, New York: Harper and Row, 1974.

[2] Eric J. Dammann, " "The Myth of Mental Illness:" Continuing controversies and their implications for mental health professionals" *Clinical Psychology Review* 17, no. 7 (1997), 738.

## 2 Szasz's ideas and their criticisms

### 2.1 The illegitimacy of mental illness as a category

Szasz's primary interest is the exclusivity of the concept of illness to physical conditions[3]. He believes that illnesses are fundamentally bodily, such that calling a mental condition an 'illness' is necessarily metaphorical. Illness requires a physiological deviation like a lesion[4]. Mental 'illnesses' are also deviations, but not from anatomical and physiological norms. Instead, Szasz considers them psychosocial and ethical deviations, since they lack the physiological deviations that constitute illness.

Several critics believe that these ideas are, essentially, Szasz (unscientifically) 'raising the spectre of dualism'[5], by discriminating mind and body. In response, Szasz explicitly denies Cartesian dualism[6]. The problem is not dualism itself; Szasz would not deny that physical illness can also cause mental suffering. Rather, he denies that mental conditions in particular should similarly fall under both categories simultaneously. If a so-called mental illness turns out to have a physiological correlate, he considers it a physical illness like any other, eliminating the concept of mental illness. Hence, Szasz willingly concedes that physiological correlates would legitimise mental conditions as illness[7].

The real problem with this argument, according to one of Szasz's most prolific critics, R.E. Kendell, is that mental features are frequently considered essential features of *bodily* illness. *No* illness is purely physical, because no illness acknowledges the mind-body distinction. Pain and suffering are as characteristic of somatic diseases as mental ones. Kendell holds that it is neither minds nor bodies but *people* who become ill[8] [9]; the prevalent conception of disease is in terms of suffering and functional impairment, transcending mind-body dualism. Szasz's conception is naïve and unfaithful to how the concept of disease has always been used.

Surprisingly, Kendell does not use this to defend psychiatry's legitimacy as equal to medicine's – he goes the other way, and inflates Szasz's stance to say that physical illness, in these terms, is just as meaningless and mythical a concept as mental illness[10] [11]. Though perhaps used as a *reductio* to demonstrate the unfeasibility or triviality of Szasz's position, I consider this possibility sincerely in Section 6.

### 2.2 'Problems in living'

A keen observer of legal and civil implications of psychiatric beliefs, Szasz did intend *methodological* dualism[12], believing that bodily disease and mental suffering should be dealt with differently. This arose from his belief that the subject-matter of mental illness is moral, not medical, because it refers to 'problems in living' that naturally populate the human condition. Psychiatry's pathologisation of these problems is the 'institutionalised denial of the tragic nature of human life'[13].

---

[3] Ibid., 734.

[4] Szasz, *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*.

[5] Dammann, " "The Myth of Mental Illness:" Continuing controversies and their implications for mental health professionals," 737.

[6] Brendan D. Kelly et al., "The Myth of Mental Illness: 50 years after publication: What does it mean today?" *Irish Journal of Psychological Medicine* 27, no. 1 (2010), 36.

[7] Hannah Pickard, "Mental illness is indeed a myth," in *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, ed. Matthew Broome and Lisa Bortolotti, New York: Oxford University Press USA, 2009, 85.

[8] Robert E. Kendell, "The nature of psychiatric disorders," in *Companion to Psychiatric Studies*, ed. Robert E. Kendell and Andrew K. Zealley, Edinburgh and London: Churchill Livingstone, 1993.

[9] Robert E. Kendell, "The Myth of Mental Illness," in *Szasz Under Fire: A Psychiatric Abolitionist Faces His Critics*, ed. Jeffrey A. Schaler, Chicago, Open Court, 2004, 40-42.

[10] Ibid., 41.

[11] Mark Cresswell, "Szasz and His Interlocutors: Reconsidering Thomas Szasz's "Myth of Mental Illness" Thesis" *Journal for the Theory of Social Behaviour* 38, no. 1 (2008), 38.

[12] Kelly et al., "The Myth of Mental Illness: 50 years after publication: What does it mean today?", 41.

[13] Thomas S. Szasz, "Diagnoses are not diseases" *The Lancet* 338, no. 8782-8783 (1991).

Szasz believed that psychiatry 'thingifies' people, treating mentally ill people like defective machines[14], and 'presumes them to be incompetent'[15]. The worry is that literal interpretation of an intended metaphor downplays the moral nature of conduct: someone's behaviour is a deviation from ethical norms so significant that it is *reminiscent* of the deviations in illness – 'they acted *as if* they were mentally ill'[16].

However, Szasz also espouses some extreme views[17]: he holds that people are *always* responsible for their behaviour, with no acts being involuntary. He also does not consider the treatment of problems in living to be the unqualified business of psychiatry or medicine. These positions are increasingly indefensible with the progress of psychiatry, but seem to invariably arise from complete commitment to methodological dualism, on which moral issues should never be approached physically. My proposals will, therefore, seek grounds on which conditions we call mental illnesses can still sometimes or somewhat be helped by interventions we call psychiatric treatments, while nonetheless denying that we should (thus) distinguish mental illness as a category.

## 3 Pickard's 'tracking' tactic

Hanna Pickard aims to vindicate Szasz by entirely circumventing the debate about the meaning of illness and the status of mental conditions as valid scientific kinds. She hypothesises the discovery of a neurophysiological correlate for schizophrenia reliable enough to make it the basis for diagnosis, much like many bodily illnesses. Upon then finding a subject who has the 'schizophrenia lesion' but lacks any schizophrenic symptoms, we would be inclined to say that the subject has schizophrenia, perhaps of a 'latent' kind as opposed to a 'full-blown' kind. But we would not intuit that the subject is *mentally ill*, and she does not deviate from 'psychosocial and ethical norms'. The argument is that the concept of 'mental illness' tracks superficial symptoms, as opposed to underlying scientific properties[18]. By 'underlying scientific properties', Pickard refers to material constituents of a condition, meaning that the tactic is to claim that what we think of as mental illness comes apart from what we deem valid scientific kinds. This tactic appears useful to my aim, since delegitimising mental illness as a category will inevitably, and as a minimum, require what Pickard calls 'scientific validity' (that is, physiological correlates) to not automatically provide indisputable basis for formal categorisation.

The caveat is that Pickard's tracking selection is vulnerable to Kendell's objection to Szasz's alleged dualism. She claims that mental illnesses track deviations from ethical norms, but also includes psychosocial deviations, referring to 'superficial or personal-level symptoms' like 'mental distress', which presumably includes mental pain and suffering. This is the opening for Kendell's view to delegitimise Pickard's basis for distinguishing mental illness as tracking superficial symptoms rather than underlying scientific properties. The superficial symptoms are an essential feature of the consensus nature of illness, which Pickard does not contest because of her intention to evade that debate. Yet, she tracks a feature that does not escape that debate, so her view collapses the distinction between physical and mental illness just as Kendell predicts. So, a tighter conception of what mental illness tracks is needed.

## 4 'Moral deviation'

This section provides a novel moral framework for Szasz's envisioned 'moral character' of mental illness. The definition outlined is to be slotted in to Pickard's 'tracking' tactic to improve its viability.

[14] Thomas S. Szasz, *Ideology and Insanity: Essays on the Psychiatric Dehumanization of Man* (London: Morion Boyars Publishers, 1973).

[15] Szasz, *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*.

[16] Theodore R. Sarbin, "On the futility of the proposition that some people be labeled "mentally ill" " *Journal of Consulting Psychology* 31, no. 5 (1967).

[17] Thomas S. Szasz, *Law, Liberty and Psychiatry: An Inquiry into the Social Uses of Mental Health Practices*, London: Routledge and Kegan Paul, 1963.

[18] Pickard, "Mental illness is indeed a myth", 87.

In the Greek tradition, eudaimonia, understood as 'flourishing' or 'wellbeing', was considered the highest human good and the goal of human life. The latter aspect is salient in Aristotle's thought: flourishing is the purpose of life, fulfilled through living in accordance with virtue[19]. Building off of this, Aristotle places the so-called pathological on a continuum with the normal, whereby possession of mental 'health' is virtue, and possession of mental 'illness' is vice, intended as inflationism ('mental illness' is just an instantiation of vice)[20].

Additional to this is ethical egoism, which states that people should pursue their own welfare[21]. This grounds the good in what is best for oneself as per the criteria determined by the human condition. Whether Aristotle himself assumed or intended egoism is controversial[22]; those who believe he did not can take me to be coopting egoism into eudaimonia for my view.

The views outlined, together, more clearly capture Szasz's belief that mental illnesses are moral problems[23]. The deviancy of the crazed, psychotic man who kills his ex-wife is easily identified on conventional morality. But it seems unreasonable on the lay view to recognise a moral deviation in the depressed or the traumatised, whose conditions can cause suffering restricted to themselves. That view only becomes sensible once 'moral deviations' mean phenomena against the wellbeing of oneself. Hereon, 'moral deviation' means 'anti-flourishing mode of being'.

## 5 The moral subject-matter of mental illness

The proposal, now, is this: all problematic mental conditions, those currently considered illnesses and those not, are united under the underlying category of moral deviations. This is the category the concept of mental illness tracks; 'mental illness' is not a distinct, unique category. Replicating Pickard's aim, this bypasses the debate about the scientific validity of mental illness, the answer to which is a strictly explanatory addition (of considerable utility, to be sure) to *some* items in the moral dimension. If mental illnesses turn out to be valid scientific kinds, we have a more detailed explanation of the cause (and possibly, treatment) of the moral deviation. But this does not affect the account of what these problems are, foremost.

The remainder of the argument is explained through examples of major depressive disorder (MDD) and generalised anxiety disorder (GAD), selected for their prevalence, but also because they may be relatively difficult, otherwise, to term moral problems.

MDD has been defined behaviourally, neurophysiologically and even phenomenologically[24]. All of these are equally tangential to this subject; they only scientifically characterise the nature of depression insofar as it exists as such a category. The idea here is that there is a morally relevant biconditional for problems in living, and what the construct of depression is tracking, scientific or otherwise, is a subset of the latter. The reader might interpret my claim to simply be that mental illnesses are functional kinds. My intention, though, is to allow a condition to exist simultaneously on material and moral accounts, so that it can *additionally* be a scientifically valid kind; the claim, however, is that the identification of the condition as a construct *at all*, *in any regard*, is in virtue of its partaking of the moral dimension – the moral aspect has priority over the physical because it is what informs the formation of the category as a problem.

[19] *Nicomachean Ethics* 1098a16.

[20] Edward Harcourt, "Aristotle, Plato and the anti-psychiatrists: Comment on Irwin," in *The Oxford Handbook of Philosophy and Psychiatry*, ed. Fulford et al., New York: Oxford University Press, USA, 2013, 47.

[21] Robert Shaver, "Egoism," *Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), accessed November 28, 2023, https://plato.stanford.edu/archives/spr2023/entries/egoism/.

[22] Tom P. Angier, "Aristotle and the Charge of Egoism" *The Journal of Value Inquiry* 52, no. 4 (2018).

[23] The remainder of this paper does not rely on the endorsement of a view as esoteric as egoism; it simply lends itself well to the point being made. Similar conclusions can be extracted on branches of virtue ethics, like Aristotle's, which identify a state of mental wellbeing with the right mode of existence, or normative theories that require the upkeep of oneself as an end rather than a means. Fundamentally, the improperness of the condition must not be grounded in contingent consequences like harm caused to the subject's loved ones, because that has reduced scope and tracking reliability.

[24] Cecily M. Whiteley, "Depression as a Disorder of Consciousness" *The British Journal for the Philosophy of Science* (2021).

So, depression is caused by the occurrence of certain personal events that disturb flourishing *such that* symptoms like low mood and self-esteem, or the phenomenological change described by Whiteley, are produced. It is not just because suffering offends the human purpose of flourishing that the depressed are morally deviant; their suffering itself is evidence of the occurrence of moral deviations in their history. Here is a development in the point made in Section 2.2. Depression does not inexplicably emerge, suddenly victimising a subject such that observers can only sympathise with her misfortune for having contracted it. It carries an extensive causal history, but is only considered of clinical interest when the buildup of this process crosses a diagnostic threshold. Such a distinction fails to appreciate the core of the condition. Depression is the consistent departure from what promotes the anti-thesis of a (clinically relevant) depressive state, such this state is sanctioned by the mind[25].

Therefore, the DSM definition of mental illness is incorrect to differentiate 'proportionate' responses sanctioned by environmental factors and seek a dysfunction sourced in the individual. What psychiatry retroactively deems a dysfunction is just as proportionate a manifestation as any other. Every other problem in living possesses a symmetrically robust and available explanation. The gap is only in human ability and willingness to articulate the explanation. An omniscient examiner of how the depressed have come to be depressed would be more surprised if they did *not* culminate in precisely their present condition.

Next, consider two people, Robert and Evan. Robert is a chronically lazy person, who wastes away and refuses to develop himself out of aversion to exertion. Evan is a diagnosed patient of GAD, who has comparable behavioural maladaptations but exhibits them out of a debilitating worry for failure and change, typical of his cognitive patterns. We have superior epistemic coverage of Evan's case by virtue of the clinical interest we take in it, while Robert's is relegated to more abstract and informal self-help remarks. However, both suffer moral problems.

Robert and Evan are similar in that they are both in a state of character that is against their flourishing, but also in that these states are evidence of a historical departure from what would have made their life go well. Evan's symptoms, whether cognitive patterns like repetitive worrying and feeling overwhelmed, or emotions like fear, insofar as they are expressive of certain damaging prior experiences he had, are akin to Robert, whose sloth is a reflection of some other feature of his character. They both have reasons, and these reasons are problems not with particular mental configurations (which are symptoms, not causes) but concerning what is good of and for a person.

It might be asked why we cannot go the other direction and say Robert has an undiagnosed form of anxiety, or something similar. However, the present line of reasoning should skew us in this direction. If it is accepted that most problems in living are fundamentally orchestrations of mundane and non-pathological events, it is naturally contrived to package and promote each of them with labels that imply pathology or distinct categories of dysfunction. Similarly, it is not that non-material psychiatric interventions like talk therapy cannot ever help subjects. We should think, rather, that to the extent they are efficacious, they are doing in essence what self-help media does for Robert, just in a more sophisticated and systemised manner.

It is also not that selective systemisation of conditions betrays the fact that some conditions uniquely benefit from it in a way that qualifies them as meaningfully distinct in *nature* to moral problems. It is simply that some moral problems have a perceived degree of complication (for example, pattern-following cognitive symptoms) that are thought to require a corresponding degree of sophistication and organisation in interventions attempting to remedy them[26]. But as has been argued, this difference in degree does not qualify as dysfunction; mental symptoms must not be conflated with the underlying problem. Clearly, then, it must also be answered why we should uphold this conception of mental symptoms as mere instrumental details in what are essentially moral problems, while not upholding the same reductive view for physical symptoms in bodily afflictions. The following sections address this.

[25] Insofar as it is not the product of a neurochemical issue – I address this in section 6 and in objection 1 in section 7.

[26] Though I cannot undertake it here, there is also an additional possible argument here that this degree of systemisation and formalised intervention is less necessary to remedy mental illnesses than physical illnesses, in that medicine is largely the only possible cure for somatic diseases, while psychiatry is much less decidedly the sole solution for mental suffering.

## 6 Explanatory construction

I now address the interaction of Kendell's objection with my view, explain why it avoids Pickard's trouble, and clarify Szasz and Kendell's positions to shed new light on the debate.

I have given my tracking slot to moral deviation instead of suffering, since suffering is a feature of all illness and hence not grounds for eliminating only mental illness. However, this move initially appears to have achieved nothing. The reason is that, especially given my 'anti-flourishing' concept of morality, a conditional is obtained with suffering as the antecedent and moral deviation as the consequent. If suffering always invokes morality, and suffering is as characteristic of bodily illness as it is of mental illness, then we apparently fail again to differentiate between the two.

My response is to bite the bullet, because, as I now show, the 'suffering' matter is something that all parties involved must and *do* concede, but it is not what my or Szasz's argument needs to rest on – only Pickard falls victim.

Szasz never denied that physical illnesses also constitute problems in living (see Section 2.1). He understood that the concept of illness, even if strictly physical, implicates moral deviation, since it involves judgement that suffering is bad for oneself. His point was not metaphysical but epistemic: 'although the desirability of physical health, as such, is an ethical norm, what health is can be stated in anatomical and physiological terms'[27]. What he meant is that physical deviations form an intelligible and useful category uniting some members of the set of moral deviations. If the connection to physicality is relaxed, the reference of illness becomes interchangeable with generic problems, and the word loses its meaning. *That* is why he restricted illness to physicality. Szasz really was not a substance dualist; he was motivated by methodological dualism, believing that it serves mental sufferers better to interpret their condition as moral rather than illness. That this was his intention is evidenced in his simple reply to Kendell: 'I disagree. The "concept of physical illness" demarcates a category… Every concept or idea can be used or abused, help people or harm people.'[28]

So, Szasz intended what I argued in Section 5 about non-moral facts being explanatory constructs, differing only in denying that they can be relevant to treating mental conditions. Bizarrely, Kendell expressly agrees about explanatory construction: 'For most of human history, disease has been essentially an explanatory concept, invoked to account for suffering'[29].

The reason Kendell did not then reach the same conclusion as Szasz was because he took mental features to be explanatorily significant. Suffering has neural correlates, which no one in this debate would deny – none are dualists. But Kendell thinks suffering only explains the problem when construed in its capacity as a mental feature. The DSM thinks likewise: it is not bold enough to adopt dualism, yet asserts psychological processes as essential constituents of mental illnesses alongside biological processes, not acknowledging that the former is ostensibly reducible to the latter[30].

So, no one is arguing metaphysics here – the disagreement is about our epistemic decisions. Finally, my proposal vindicates Szasz's intention: having assumed illness to be an explanatory construct, I claim that moral features are *better explanations* of suffering than mental features (which is the departure from Kendell) – and I have already argued for precisely this in Section 5. Additionally, we preserve physical illness; (using Szasz's point) physical explanations are still useful to explaining *some* moral deviations (see Objection 1 for clarification). Consequently, mental illness alone is eliminated.

The explanatory reduction of the psychological to the physical is demonstrable in any disorder where phenomenological or cognitive symptoms are caused by lesions. It goes through straightforwardly because the opposition also promotes physical explanations; I am seeking to replace

[27] Szasz, *Law, Liberty and Psychiatry: An Inquiry into the Social Uses of Mental Health Practices*.

[28] Thomas S. Szasz, "Reply to Kendell," in *Szasz Under Fire: A Psychiatric Abolitionist Faces His Critics*, ed. Jeffrey A. Schaler, Chicago, Open Court, 2004, 54.

[29] Kendell, "The Myth of Mental illness," 31.

[30] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (Washington: American Psychiatric Publishing, 2013).

only the mental (with the moral). That reduction goes through because Kendell was incorrect to think that it is the mental formulation of suffering that makes it relevant. It is apprehended as a mental phenomenon, but only qualifies as a relevant criterion by virtue of moral deviation being its consequent.

Also, as argued in Section 5, mental deviations in conditions like depression or anxiety are often eventualities orchestrated by a massive sequence of micro- or macro-events related to the human good. Where they are instead closer to having apparently 'dropped from the sky', the explanation is likely to be biological/neurophysiological, invoking the first category. The idea is that physical and moral features are *jointly* sufficient to eliminate mental features as meaningful representatives of moral deviation, such that mental features are always symptomatic of deviations in another category.

## 7 Objections and replies

*Objection 1: Even if the mental to physical explanatory reduction goes through, the mental to moral move is dubious. Mental features like emotions and beliefs, barring ones symptomatic of neurophysiological defects, directly generate 'anti-flourishing', and hence the mental category remains the best explanation of the condition.*

The answer to this this was teased out in Section 5, but bears restating. It is true that somatic and neurophysiological, as well as cognitive (assuming suspension of the notion that cognition is reducible to neurophysiology anyway) deviations, can all create moral deviations. The argument, however, is not targeting contribution to anti-flourishing, but being manifestative of anti-flourishing. In this regard, physical afflictions are distinct from mental ones.

Bodily illnesses warrant acknowledging and isolating as an explanatory category because they are 'starting points': a lesion can cause suffering (→moral deviations), but its acquisition is unrelated to the moral dimension, and arbitrary in that sense. Meanwhile, mental features that constitute suffering, such as those in mental illnesses, cause moral deviations but are themselves existent due to a prior moral deviation. Because they are reflective of moral features, their subject-matter is best explained and understood as a straightforward moral problem.

Compare Evan's GAD to an athlete who contracts a disease that sidelines him, damaging his career. Evan's suffering is reflective of something that happened to him that was against the human good; perhaps he was bullied or abused in childhood. The mental features are only responses sanctioned by the experience of suboptimal life events. The athlete, meanwhile, experiences moral disruption due to something that cannot be explained morally in the first instance (changes to his body).

Of course, GAD potentially has a genetic component, just as several mental afflictions are potentially influenced by neurochemical issues. That is why the reasoning for the elimination of mental features as an original explanation is an *inclusive disjunction* of physical and moral features. This is also where my view is advantageous over Szasz's *exclusive* disjunction, which does not allow for a condition to be explained partly physically and partly morally – we are reluctant to treat depression, for example, as a mere brain disorder like dementia, but do not want to rule out material influences on it.

*Objection 2: In Section 5, you criticised psychiatry's practice of labelling moral problems 'illnesses' simply because they outwardly crossed a diagnostic threshold. However, this practice is commonplace in medicine for conditions that are clearly illnesses, like diabetes.*

Because, as Kendell notes, illness is a pragmatic construct, the crossing of a quantitative threshold rather than the undergoing of a qualitative change is indeed a staple criterion of medical diagnoses. However, the difference is that in medical cases, the phenomenon only begins infringing on one's wellbeing *after* crossing the diagnostic threshold, which is why the threshold's placement is legitimate. In the case of depression that I argued for in Section 5, the threshold is crossed in the first place *because* flourishing was disturbed. Therefore, illness (insofar as that term alludes to a physically observable metric) is the better explanation for diabetes, but morality remains the better

explanation for depression (as opposed to mental features; to reiterate, I do allow physical features as a coexistent and meaningfully distinct explanation).

*Objection 3: Unlike Szasz, you concede that the subject-matter of 'mental illnesses' can be, partially, of medical relevance. So, your conclusion is trivial, or simply a pedantic reflection on human suffering.*

The progress of psychiatry has made it simply incorrect to claim that medical interventions are never of use. The hope with this paper is not to dispose of all psychiatric interventions for problems in living, but to clarify how problems are best understood, and hence, treated. Meaning, I retain Szasz's methodological dualism only partially, because full commitment generates untenable positions.

My view avoids Szasz's conclusion that people are always responsible for their conduct. To the extent that a moral deviation is better explained physically, we can absolve individuals of responsibility. Of course, it is still a moral deviation – morality and agency come apart on the 'anti-flourishing' conception, which tracks the set of moral patients, not the set of moral agents.

However, mental conditions explained *more* as moral deviations than physical deviations should be interpreted *primarily* as indications that the subject's life is not conducive to her well-being, rather than pinning the suffering on a disorder, an additional entity. People have excellent reasons for being depressed, anxious or traumatised, and eagerness to alleviate behavioural symptoms/manifestations (due to perception of the issue as consisting in mental symptoms rather than a deeper moral problem) via medication or even therapeutic intervention may sometimes overlook the root cause. Any 'dysfunction' spoken of should be metaphorical, and with respect to the patient's life.

## 8 Conclusion

Szasz argued that mental illnesses constitute moral problems and not scientific constructs like disease. He was pegged as a substance dualist illegitimately discriminating against mental illness. I have combined Pickard's 'tracking' tactic with a novel conception of morality to argue in defence of Szasz's intention. I have claimed that all illness is manifestative of moral deviation *qua* suffering, but that distinctions are possible on epistemic grounds. Mental features are considered essential to the explanation of illness, but they are better explained in either physical or moral terms, such that the only categories required are 'illness', which is physical, and moral deviations. The former remains a subset of the latter, allowing a departure from Szasz's strict methodological dualism, which invited the criticisms he drew. Barring that, Szasz is misunderstood, and quietly insightful in claiming that mental illness is a myth.

## Bibliography

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*, 5th ed. Washington: American Psychiatric Publishing, 2013.

Angier, Tom P. "Aristotle and the Charge of Egoism." *The Journal of Value Inquiry* 52, no. 4 (2018), 457-475. doi:10.1007/s10790-018-9632-2.

Aristotle. *Nicomachean Ethics*.

Cresswell, Mark. "Szasz and His Interlocutors: Reconsidering Thomas Szasz's "Myth of Mental Illness" Thesis." *Journal for the Theory of Social Behaviour* 38, no. 1 (2008), 23-44. doi:10.1111/j.1468-5914.2008.00359.x.

Dammann, Eric J. ""The Myth of Mental Illness:" Continuing controversies and their implications for mental health professionals." *Clinical Psychology Review* 17, no. 7 (1997), 733-756. doi:10.1016/s0272-7358(97)00030-5.

Harcourt, Edward. "Aristotle, Plato and the anti-psychiatrists: Comment on Irwin." In *The Oxford Handbook of Philosophy and Psychiatry*, edited by Bill Fulford, Martin Davies, Richard Gipps, George Graham, John Z. Sadler, Giovanni Stanghellini, and Tim Thornton, 47-52. New York: Oxford University Press, 2013.

Kelly, Brendan D., Pat Bracken, Harry Cavendish, Niall Crumlish, Seamus MacSuibhne, Thomas S. Szasz, and Tim Thornton. "The Myth of Mental Illness: 50 years after publication: What does it mean today?" *Irish Journal of Psychological Medicine* 27, no. 1 (2010), 35-43. doi:10.1017/s0790966700000902.

Kendell, Robert E. "The nature of psychiatric disorders." In *Companion to Psychiatric Studies*, edited by Robert E. Kendell and Andrew K. Zealley, 1-8. Edinburgh and London: Churchill Livingstone, 1993.

——. "The Myth of Mental Illness." In *Szasz Under Fire: A Psychiatric Abolitionist Faces His Critics*, edited by Jeffrey A. Schaler, 29-48. Chicago: Open Court, 2004.

Pickard, Hanna. "Mental illness is indeed a myth." In *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, edited by Matthew Broome and Lisa Bortolotti, 83-102. New York: Oxford University Press, USA, 2009.

Sarbin, Theodore R. "On the futility of the proposition that some people be labeled "mentally ill."." *Journal of Consulting Psychology* 31, no. 5 (1967), 447-453. doi:10.1037/h0025018.

Shaver, Robert. "Egoism." Stanford Encyclopedia of Philosophy (Spring 2023 Edition). Accessed November 28, 2023. https://plato.stanford.edu/archives/spr2023/entries/egoism/.

Szasz, Thomas. "Diagnoses are not diseases." *The Lancet* 338, no. 8782-8783 (1991), 1574-1576. doi:10.1016/0140-6736(91)92387-h.

——. *Ideology and Insanity: Essays on the Psychiatric Dehumanization of Man*. London: Marion Boyars Publishers, 1973.

——. *Law, Liberty and Psychiatry: An Inquiry into the Social Uses of Mental Health Practices*. London: Routledge and Kegan Paul, 1963.

——. *The Myth of Mental Illness: Foundations of a Theory of Personal Conduct*, 2nd ed. New York: Harper & Row, 1974.

——. "Reply to Kendell." In *Szasz Under Fire: A Psychiatric Abolitionist Faces His Critics*, edited by Jeffrey A. Schaler, 49-56. Chicago: Open Court, 2004.

Whiteley, Cecily M. "Depression as a Disorder of Consciousness." *The British Journal for the Philosophy of Science*, 2021. doi:10.1086/716838.

# When Self-Trust and Peer-Trust Collide

Justin Lee, *University of St Andrews*

According to the Asymmetry View, one rationally ought to have more epistemic self-trust than trust in one's disagreeing epistemic peer unless one has case-specific reasons not to (e.g. one is drunk during the given disagreement). In this essay, I argue that the Asymmetry View is wrong as a general principle of how to balance epistemic self-trust and trust in one's peer. To this end, I challenge Enoch's argument from the ineliminability of the first-person perspective, which I deem the most compelling defence of this principle. I concede that Enoch could defend a more modest version of the Asymmetry View by altering his argument to account for my criticisms. Nonetheless, I stress that this modified principle is applicable only in rare and indeed unrealistic cases.

## 1 Introduction

My buddy Solomon and I are enjoying our neighbourhood café's new chocolate mousse. Judging by its flavour profile and what I know about various couvertures on the market, I think it is made from Guanaja 70%. Solomon disagrees; Caraïbe 66% is his conclusion. Through our past conversations, though, we have come to consider ourselves epistemic peers on chocolate-related matters. That is, we each deem the other equally well informed about chocolate and equally reliable in judging chocolate-related information. Should I therefore maintain or lower confidence in my belief about the mousse, or perhaps suspend judgement? This touches the core of recent philosophical debates about disagreement: what is the most epistemically rational, i.e., evidentially supported and logical, response to a doxastic disagreement with one's supposed epistemic peer?

A crucial factor in answering this is the balance one should have between epistemic self-trust and trust in one's peer (hereon "peer-trust"). When locked in disagreement with Solomon, is it epistemically rational for me to trust my epistemic faculties—senses, inferential capacities, etc.—more than his? The Symmetry View says it is not—I should have equal self-trust and peer-trust. This supports lowering confidence or suspending judgement. The Asymmetry View (hereon "AV"), meanwhile, says I should have more self-trust than peer-trust unless I have case-specific reasons not to (hereon "specific defeaters"). A possible specific defeater is that I am drunk while Solomon is not, since this means his faculties are probably more reliable on this occasion. AV thus supports maintaining confidence, absent such a specific defeater.

In this essay, I argue that AV is wrong as a general principle of how to balance epistemic self-trust and peer-trust.[1] Except in some rare and unrealistic cases, it is not epistemically rational to have more self-trust than peer-trust, even absent specific defeaters.

I begin by explicating what I deem the most compelling argument for AV: Enoch's argument from the ineliminability of the first-person perspective (hereon "1PP"). I then turn to Peter's objection that one of Enoch's premises is misleading and consider how Enoch might respond by distinguishing between two types of epistemic rationality. Thereafter, I maintain that, even with this distinction in place, said premise remains in trouble because it ignores two general defeaters, i.e., defeaters that arise in virtually all peer disagreements: (i) our recognition of our own epistemic fallibility and (ii) our appreciation of higher-order symmetry. This means AV, as a general principle, must fall. I concede that Enoch could defend a more modest version of AV by altering the offending

[1] While this undermines a crucial source of support for the option of maintaining confidence, I do not address whether we should ultimately reject said option. My focus is squarely on the balance of trust during peer disagreements.

premise to account for these general defeaters. Nonetheless, I stress that this modified principle is applicable only in rare and indeed unrealistic cases.

## 2 Enoch: Argument for AV

Enoch begins by observing that my 1PP has an "ineliminable role" in the "form[ation] and revisi[on] [of my] beliefs".[2] That is, to determine what to believe, I must rely on how things seem to *my* epistemic faculties. This makes epistemic self-trust necessary for getting me anywhere in my epistemic life.[3]

Among my beliefs, Enoch points out, are those about the reliability of others.[4] [5] Even in judgements I make of another's reliability, and thus of how much to trust them, then, my 1PP is ineliminable. To determine how much peer-trust to have in Solomon, I must rely on *my* faculties to assess how reliable his are.

But *how* do my faculties assess Solomon's? Enoch says it surely has to heavily involve how often *my* faculties consider *his* correct on chocolate-related matters, i.e., on his epistemic track record regarding chocolate *as I see it* (or *as it seems from my 1PP*).[6] Each time he is right *as I see it*, I gain evidence that he is as reliable as I am.[7] Each time he is wrong *as I see it*, I gain evidence that he is less reliable than I am.

Given this evidential situation, Enoch concludes that, absent specific defeaters, whenever one disagrees with someone they initially regard as an epistemic peer, one should come to trust said peer less in light of that very disagreement.[8] I have, drawing on past conversations, judged Solomon to be equally reliable. Hence, prior to our disagreement, I should have equal self-trust and peer-trust. However, I now deem him wrong about the mousse. This makes it epistemically rational for me to demote him from peerhood, i.e., to trust his faculties less than mine.[9]

For clarity, Enoch's argument may be rendered thus:

> P1) My 1PP is ineliminable from my assessment of my peer's reliability.
>
> P2) If P1, then a significant part of my evidence for my peer's reliability is their track record *as I see it*.
>
> SC) A significant part of my evidence for my peer's reliability is their track record *as I see it*. [from P1 and P2]
>
> P3) If SC, then absent specific defeaters, it is epistemically rational for me to have more epistemic self-trust than peer-trust during a disagreement.
>
> C) Absent specific defeaters, it is epistemically rational for me to have more epistemic self-trust than peer-trust during a disagreement. [from SC and P3]

## 3 Peter: Track Record Misleads

In response to Enoch, Peter holds that using my peer's track record in the way he describes would mislead me, for it neglects the possibility that the disagreement could be the result of my, and

[2] Enoch, "Not Just a Truthometer", 962.

[3] Enoch, 980.

[4] Enoch, 973.

[5] Rattan, "Disagreement and the First-Person Perspective", 36.

[6] Enoch, 973.

[7] So, for Enoch, Solomon's being right *as I see it* would not offer me evidence that he is *more* reliable than I am. Absent specific defeaters, then, my self-trust would, as Schafer puts it, "constrain" my peer-trust; see Schafer, "How Common Is Peer Disagreement?", 31.

[8] Enoch, 974, 980.

[9] This procedure for determining how to balance self-trust and peer-trust may immediately strike the reader as question begging. See Enoch, 980-1, for his reply that such question begging is unproblematic and indeed necessary to prevent radical scepticism. I find this reply suspicious, but do not contest it in what follows, since radical scepticism is beyond the scope of this essay.

not Solomon's, getting things wrong.[10] Since I have antecedently judged him to be my peer on chocolates, I should deem him no more likely than I am to err on a chocolate related-matter. When the disagreement about the mousse arises, then, it should be unclear to me which one of us has arrived at the wrong conclusion. I cannot simply assume he is the one who is wrong just because he is wrong *as I see it*. Hence, I do not actually have reason to trust him less.

If Peter is correct, then P3 is misleading and false. That a significant part of my evidence for Solomon's reliability is his track record *as I see it* does not make it epistemically rational for me to trust him less than myself when we disagree. Such a demotion, Peter emphasises, would make sense only if I can eliminate the possibility that my opinion about the mousse is wrong, and thus be assured that it is Solomon who has erred.[11]

## 4 Enoch: First-Person VS Third-Person Epistemic Rationality

Enoch, however, might reply that Peter fails to appreciate the evidential situation of my 1PP as party to the disagreement. My epistemic faculties have processed direct evidence—sensory inputs and relevant background knowledge—about the mousse and concluded that it is made from Guanaja.[12] From my 1PP, then, Solomon is wrong. Crucially, Enoch acknowledges the possibility that, as a matter of fact, Solomon is not wrong[13]—this is the heart of Peter's objection. However, he insists that, since the direct evidence *as seen from my 1PP* supports the conclusion that Solomon is wrong, this conclusion is legitimate evidence *for my 1PP* that he is less reliable than I am.[14] Thus, Enoch maintains that I am epistemically rational in having more self-trust than peer-trust. For clarity, he could specify that I am *first-person* epistemically rational in so balancing my trust, a balance of trust being rational in this sense if it is supported by the evidence *as seen from the trust giver's 1PP*.

Contrastingly, Enoch might hold, Peter's picture of the disagreement presupposes the evidential situation of one who takes the third-person perspective (hereon "3PP") to it.[15] Suppose Peter consults Solomon and me on the mousse. She deems us equally well informed and reliable, and thus deserving of equal trust on the matter. I say it is made from Guanaja and Solomon denies this. At least one of us must be wrong, but she has no evidence as to what the mousse is made of other than her prior evidence that Solomon and I deserve equal trust and our currently conflicting testimonies.[16] Since she has no direct evidence about the mousse like we do, it is unclear to her which one of us has erred.[17] Thus, an explanation of the disagreement in terms of Solomon's being less reliable is, as seen from her 3PP, no more reasonable than one in terms of my being less reliable. This explains why my having more self-trust than peer-trust looks like the result of my being misled, and thus irrational, from her 3PP. Indeed, Enoch could specify that it is admittedly not *third-person* epistemically rational for me to have more self-trust than peer-trust because *the evidential situation of the 3PP* to the disagreement does not support this balance.

Hence, Peter, as occupant of the 3PP, can charge me with irrationality. I might even appreciate from my 1PP that she could do so from her 3PP, and thus understand that it would be third-person irrational for me to have more self-trust than peer-trust.[18] However, Enoch would insist, it is, in a different sense, epistemically rational for me, as occupant of a 1PP in the disagreement, to have such a balance of trust, since this is what my 1PP's evidential situation supports. I can be first-person

[10] Peter, "Epistemic Self-Trust and Doxastic Disagreements", 1196.
[11] Peter, 1196.
[12] Enoch, 986, fn. 62.
[13] Enoch, 984.
[14] Enoch, 984.
[15] Enoch, 960-2.
[16] Enoch, 986, fn. 62.
[17] If she obtains direct evidence and forms a belief about the mousse based on it, she will automatically be party to the disagreement. Her 1PP might side with mine or Solomon's, or perhaps disagree with both.
[18] Enoch, 986, fn. 62.

rational while being third-person irrational.[19]

With this distinction between two types of epistemic rationality in place, Enoch might say that what he is really arguing for is not that it is epistemically rational *simpliciter* for me to have more self-trust than peer-trust given Solomon's track record *as I see it*. Rather, it is *first-person* rational. Hence, Peter's criticism, which highlights my *third-person* irrationality, poses no threat to $P_3$. For clarity, Enoch might alter $P_3$ as such:

> $P_3$*) If SC, then absent specific defeaters, it is *first-person* epistemically rational for me to have more epistemic self-trust than peer-trust during a disagreement.

In what follows, then, I take C* to be the general principle of balancing self-trust and peer-trust that Enoch defends:

> C*) Absent specific defeaters, it is *first-person* epistemically rational for me to have more epistemic self-trust than peer-trust during a disagreement.

## 5 Does the Distinction Really Protect Enoch's Argument?

Even accepting the aforementioned distinction, however, Enoch's argument is left open to Peter's criticism, for I can appreciate *from my 1PP* that I am epistemically fallible.

As noted above, Peter highlights that Solomon's being wrong *as I see it* can license my demoting him only if I can eliminate the possibility that I am wrong, and thus be assured that it is in fact Solomon who has erred. However, I, like most, can appreciate even from my 1PP that my faculties are fallible. I acknowledge that I am not fully reliable even as regards chocolate, so I have no reason to think my answer secured from error. Therefore, I should indeed consider the possibility that it is because I have erred that Solomon and I are locked in disagreement, just as Peter counsels. Moreover, since our past conversations have evidenced to my 1PP that we are equally reliable, his erring would seem no more likely than mine. My appreciation of my fallibility thus defeats my purported licence to have more self-trust than peer-trust.

To diagnose where Enoch's argument goes astray, I call attention to how Enoch states that my peer's track record *as I see it* is "a *significant* part of [my] evidence as to [his] reliability"[20] and yet seems to assume it is the *only* relevant evidence for my peer's reliability, absent specific defeaters.[21] He does not acknowledge defeating evidence other than specific defeaters like drunkenness that would be relevant to peer disagreements. However, it would be epistemically irrational for me to balance my trust without accounting for another important piece of evidence, i.e., my fallibility. This would be to privilege a subset of my total evidence, and thus be misled, as Peter maintains.

## 6 The Upshot and A Pushback

It should be underlined that recognition of one's fallibility is commonplace and thus a general defeater, i.e., a defeater that would affect virtually any peer disagreement. This should make it extremely clear that $P_3$* is false. In just about any case, recognition of one's fallibility would be present to defeat one's purported licence to demote one's peer, making such demotion first-person irrational despite the significance of the peer's track record *as one sees it*. Enoch's argument is thus unsound, and AV as a general principle of how to balance trust is undermined.

Admittedly, Enoch could push back with counterexamples. What about cases wherein an agent has no recognition of their fallibility? If I am infallible as regards chocolate, for example, no evidence of my fallibility in this domain could arise. Or perhaps, even if I am fallible, I could be truly ignorant about or have forgotten any evidence of this. If so, then it should be first-person

[19] Cf. Foley's distinction between internalist and externalist justifications; see Foley, *Intellectual Trust in Oneself and Others*, 21.

[20] Enoch, 973; emphasis added.

[21] Indeed, Peter takes Enoch to consider the track record the only relevant evidence; see Peter, 1194.

rational for me to trust myself more than Solomon, for this would be supported by evidence I actually possess—namely, his track record *as I see it*. And it would not do, Enoch might maintain, to object that being ignorant or forgetful are epistemic failings. For, as Foley observes, not every epistemic failing constitutes irrationality.[22] So long as I respond to all relevant evidence accessible to my 1PP, I am first-person rational. Thus, Enoch could propose that P3* be tweaked as such:

> P3**) If SC, then absent specific defeaters and recognition of my fallibility, it is *first-person* epistemically rational for me to have more epistemic self-trust than peer-trust during a disagreement.

This pushback seems reasonable, and while it cannot salvage AV, it at least allows Enoch to defend a modified version of it, as captured by C**:

> C**) Absent specific defeaters and recognition of my fallibility, it is *first-person* epistemically rational for me to have more epistemic self-trust than peer-trust during a disagreement.

Yet, I stress again that recognition of one's fallibility is commonplace. It is foolish to think any human could be infallible in a given domain. Moreover, aside from the highly delusional or cognitively impaired, it would be practically impossible for us who are so constantly and thoroughly fallible to remain ignorant or forgetful of this human predicament. Thus, this modified AV licenses demoting one's peer only in very rare and indeed unrealistic cases.

Unfortunately for Enoch, even P3** would not be enough to patch up his argument, for P3* faces a problem other than that which this new premise is designed to circumvent.

## 7 Christensen and Rattan: Revenge of Higher-Order Symmetry

I shall look at two ways this problem could be framed.

Christensen presents it as follows.[23] That I disagree with Solomon counts against his reliability—this aligns with P3*. However, that he—someone I have antecedently judged from my 1PP to be equally reliable—disagrees with me also counts against my reliability, and this is something I can appreciate even from my 1PP. Hence, while P3* licenses demoting Solomon in light of his track record *as I see it*, it seems, even from my 1PP, that I should demote myself too. Thus, I can appreciate a higher-order symmetry between us. Ultimately, Solomon and I should appear equally reliable to my 1PP, which defeats my having more self-trust than peer-trust, *contra* P3*.

Meanwhile, Rattan, taking inspiration from Christensen, puts the problem thus.[24] P3* licenses anyone to asymmetrically privilege their own faculties in the balance of self-trust and peer-trust, given their peer's track record. Therefore, when Solomon and I disagree, I can appeal to P3* to privilege my faculties and Solomon can make a similar appeal for his. That Solomon can do so is something I can appreciate from my 1PP, which means I can appreciate a higher-order symmetry between us. This pushes my 1PP towards according equal trust to our faculties, *contra* P3*.

### 7.1 Enoch: Higher-Order Symmetry is Irrelevant

Enoch, however, insists that higher-order symmetry is no threat to his argument.[25] When I demote Solomon in light of our disagreement, he emphasises, I am *not* demoting him on this basis: "Solomon believes it is not Guanaja, whereas *I believe it is* Guanaja", i.e., we have different beliefs. If I am, Enoch concedes, the disagreement should count equally against both parties and thereby generate higher-order symmetry. However, my 1PP's actual basis for demotion is this: "Solomon believes it is not Guanaja, whereas *it is* Guanaja", i.e., Solomon is wrong. This means

[22] Foley, 42.
[23] Christensen, "Epistemology of Disagreement", 196.
[24] Rattan, 34.
[25] Enoch, 981-3.

the disagreement in no way commits me to symmetry, forwards Enoch, for only Solomon seems wrong from my 1PP. I am therefore first-person rational in trusting my faculties more.

However, I shall contend that this reply fails for two closely related reasons.

## 8 Higher-Order Symmetry's Revenge is Not Over

First, even the asymmetry Enoch uses in attempt to fend off the problem can give rise to appreciation of higher-order symmetry. Yes, when I disagree with Solomon, I deem him wrong. But, surely, I can also appreciate from my 1PP that Solomon thinks, from his 1PP, that I am wrong. Individuals considering each other wrong is constitutive of disagreement, after all. This generates a higher-order symmetry my 1PP can appreciate. Moreover, since my 1PP has antecedently judged him to be equally reliable, an explanation of our each taking the other wrong in terms of his being less reliable should strike my 1PP as no more reasonable than one in terms of my being less reliable, *contra* P3*.

Secondly, it seems this higher-order symmetry I have just highlighted is already embedded in the aforementioned presentations of the problem, which means Enoch's reply attacks a strawman. To illustrate, let us look closer at how Christensen and Rattan frame it.

Christensen writes that "my discovering that my friend has reached what seems to me to be the *wrong* conclusion does constitute evidence that [he] has made a *mistake*" and that "the fact that [he] disagrees with [me] also constitutes evidence that I have made a *mistake*".[26] Rattan, meanwhile, writes that it initially seems I can leverage P3* because "only *my* reasoning appears *right* from my [1PP]" and that my peer can make a "*similar* invocation" of her 1PP.[27]

Both philosophers present the problem in terms of my deeming my peer wrong and myself right as well as recognition that my peer deems me wrong and himself right. Neither seems to be concerned with the mere difference in belief Enoch targets.[28]

## 9 Enoch: Can You Really Appreciate Higher-Order Symmetry?

There is, however, another response Enoch might offer. Enoch recognises that Solomon considers me wrong and himself right,[29] but nowhere acknowledges that this asymmetry, and thus the higher-order symmetry, can be appreciated from my 1PP. Perhaps Enoch might use this as the basis for a pushback. For if it turns out that one cannot appreciate the relevant higher-order symmetry from their 1PP, then it may be first-person rational for them to trust themselves more than their peer.

However, in most cases, parties to a peer disagreement would be able to appreciate that their opponent thinks they are wrong, and thus the relevant higher-order symmetry. This means such appreciation constitutes another general defeater. To illustrate this, I return to my disagreement with Solomon. I can appreciate that, from Solomon's 1PP, he thinks "this fool believes it is Guanaja, whereas *it is not* Guanaja", i.e., that I am wrong. That is why I naturally want to persuade him to think "*it is* Guanaja", i.e., that while he thinks I am wrong, I am actually right. If I can only appreciate that our beliefs are different, however, I would think that from his 1PP, he only thinks "this fool believes it is Guanaja, whereas *I believe it is not* Guanaja". But my natural urge during a disagreement is not to persuade him to think "*I believe it is* Guanaja", i.e., that while he thinks he has a view different from mine, he actually does not. I take it that, in virtually every case of peer disagreement, we are similarly disposed to persuade our peers to *believe as we believe*, and not to *believe that they actually believe as we believe*. Perhaps this means we are *naturally disposed* to trust ourselves more than our peers during disagreements with them. Crucially, though, it also evidences that we do indeed appreciate our peers' considering us wrong, and thus the relevant

[26] Christensen, 196; emphases added.
[27] Rattan, 34; last two emphases added.
[28] It seems the underlying problem is that Enoch misinterprets Christensen's argument; see bottom of Enoch, 975.
[29] Enoch, 984.

higher-order symmetry, which makes us *first-person rationally required* to have equal self-trust and peer-trust.

Admittedly, if I truly cannot appreciate that Solomon thinks I am wrong, it may indeed be first-person rational, though in another sense an epistemic blunder, for me to trust my faculties more than his. But it is hard to even imagine cases wherein one cannot appreciate that one's disagreeing peer thinks one is wrong.

## 10 Another Upshot and Another Possible Pushback

First-personal appreciation that my peer thinks I am wrong, and thus of higher-order symmetry, offers another reason to deem $P_3^*$ false. This reinforces that AV is wrong as a general principle.

Moreover, given such appreciation, even $P_3^{**}$ would not be enough to repair Enoch's argument. Absent specific defeaters and recognition of my fallibility, if I can appreciate that someone I have antecedently judged to be equally reliable thinks I am wrong, and thus appreciate the relevant higher-order symmetry between us, it would not be first-person rational to have more self-trust than peer-trust. Indeed, if I did not obtain evidence of my fallibility prior to our disagreement, that my peer thinks I am wrong now supplies such evidence, which should make me aware of said fallibility.

If Enoch wants to insist that there are still some cases wherein it is first-person rational to have more self-trust than peer-trust, he would have to tweak $P_3^{**}$ and $C^{**}$ as such:

> $P_3^{***}$) If SC, then absent specific defeaters, recognition of my fallibility, and appreciation that my peer thinks I am wrong, it is *first-person* epistemically rational for me to have more epistemic self-trust than peer-trust during a disagreement.
>
> $C^{***}$) Absent specific defeaters, recognition of my fallibility, and appreciation that my peer thinks I am wrong, it is *first-person* epistemically rational for me to have more epistemic self-trust than peer-trust during a disagreement.

I take it as uncontroversial, however, that cases wherein a party to a disagreement lacks *all* defeaters mentioned in $C^{***}$ are extremely rare and unrealistic, which foregrounds just how limited in applicability this modified AV is.

## 11 Conclusion

To conclude, AV is wrong as a general principle of how to balance epistemic self-trust and peer-trust, despite Enoch's argument from the ineliminability of one's 1PP in assessing an epistemic peer's reliability. While Enoch assumes said ineliminability entails that, absent specific defeaters, we should always deem our peers wrong, and thus deserving of less trust, in light of our disagreements, Peter rightly notes that this ignores the possibility that it is we who have erred. Though Enoch might dismiss this objection by distinguishing between first- and third-person forms of epistemic rationality, Peter's criticism still holds because of the general defeater that is our first-personal recognition of our epistemic fallibility. Moreover, our first-personal appreciation that our peers deem us wrong, and thus of higher-order symmetry, constitutes another general defeater of AV's counsel to trust oneself more than one's disagreeing peer. Moreover, while Enoch might alter one of his premises to sidestep these defeaters and defend a modified AV, this modified principle licenses having more self-trust than peer-trust only in rare and indeed unrealistic cases.

## Bibliography

Christensen, David. "Disagreement, Question-Begging and Epistemic Self-Criticism." *Philosophers' Imprint* 11 (2011). https://philarchive.org/rec/CHRDQA.

———. "Epistemology of Disagreement: The Good News." *The Philosophical Review* 116, no. 2 (2007): 187–217. https://www.jstor.org/stable/20446955.

Enoch, David. "Not Just a Truthometer: Taking Oneself Seriously (but Not Too Seriously) in Cases of Peer Disagreement." *Mind* 119, no. 476 (October 1, 2010): 953–97. https://doi.org/10.1093/mind/fzq070.

Foley, Richard. *Intellectual Trust in Oneself and Others*. Cambridge Studies in Philosophy. Cambridge: Cambridge University Press, 2001. https://doi.org/10.1017/CBO9780511498923.

Frances, Bryan, and Jonathan Matheson. "Disagreement." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2019. Metaphysics Research Lab, Stanford University, 2019. https://plato.stanford.edu/archives/win2019/entries/disagreement/.

Kappel, Klemens. "Trust and Disagreement." In *The Routledge Handbook of Trust and Philosophy*. Routledge, 2020.

McLeod, Carolyn. "Trust." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Fall 2023. Metaphysics Research Lab, Stanford University, 2023. https://plato.stanford.edu/archives/fall2023/entriesrust/.

Pasnau, Robert. "Disagreement and the Value of Self-Trust." *Philosophical Studies* 172, no. 9 (September 1, 2015): 2315–39. https://doi.org/10.1007/s11098-014-0413-x.

Peter, Fabienne. "Epistemic Self-Trust and Doxastic Disagreements." *Erkenntnis* 84, no. 6 (December 2019): 1189–1205. https://doi.org/10.1007/s10670-018-0004-x.

Rattan, Gurpreet. "Disagreement and the First-Person Perspective." *Analytic Philosophy* 55, no. 1 (2014): 31–53. https://doi.org/10.1111/phib.12038.

Schafer, Karl. "How Common Is Peer Disagreement? On Self-Trust and Rational Symmetry." *Philosophy and Phenomenological Research* 91, no. 1 (2015): 25–46. https://doi.org/10.1111/phpr.12169.

# Contributors

## Ethan Reiter

Ethan Reiter is a recent graduate of the University of Chicago, where he studied philosophy and cognitive science. His primary interests lie in consciousness and the philosophy of language, and he aims to integrate these topics with contemporary scientific research on the mind.

## Ethan Samuel Kovnat

Ethan studied philosophy and linguistics at Cornell University (having graduated in May 2024) and will be beginning a PhD in philosophy at Brown University in September. His philosophical interests lie primarily in ethics, especially issues of moral agency, responsibility, and motivation.

## Krisztian Kos

Krisztian Kos is a rising fourth year student at the University of St Andrews studying Philosophy and Physics, while reading (and occasionally writing) poetry and philosophical pieces in his free time. At present, he is engaged with issues in epistemology and philosophy of mind, and has avid interests in metaphilosophy, history of philosophy, philosophy of science and philosophy of AI.

## Rohan Mavinkurve

Rohan is a rising fourth-year Joint Honours Philosophy and Psychology student at the University of St Andrews. He is primarily interested in moral psychology and particularly enjoys Aristotlean ethics; he is also generally interested in axiological thought and how it influences individual theories of ethics and aesthetics.

## Justin Lee

Justin is a Philosophy conversion diploma student at the University of St Andrews, proceeding onto a Philosophy MLitt. His primary interests are in Epistemology and Metaphilosophy.

# *Aporia*

Undergraduate Journal of the St Andrews Philosophy Society

VOLUME XXIV

*Aporia* is funded by the University of St Andrews Philosophy Society, which receives funds from the University of St Andrews Department of Philosophy, the University of St Andrews Students' Association, and independent benefactors.
*Aporia* is published by The University of St Andrews Philosophy Society