Designed and typeset at Durham University by the Editor of the journal

Published by Durham University

ISSN 2752-7255

Critique welcomes the submission of unsolicited manuscripts of articles, reviews, discussions pieces and suggestions for symposia or special issues by undergraduates worldwide. Submission guidelines can be found at:

https://www.durham.ac.uk/departments/academic/philosophy/undergraduate-study/critique/

Address for enquiries and submissions:

durhamcritique@gmail.com

Social media:

https://twitter.com/durhamcritique

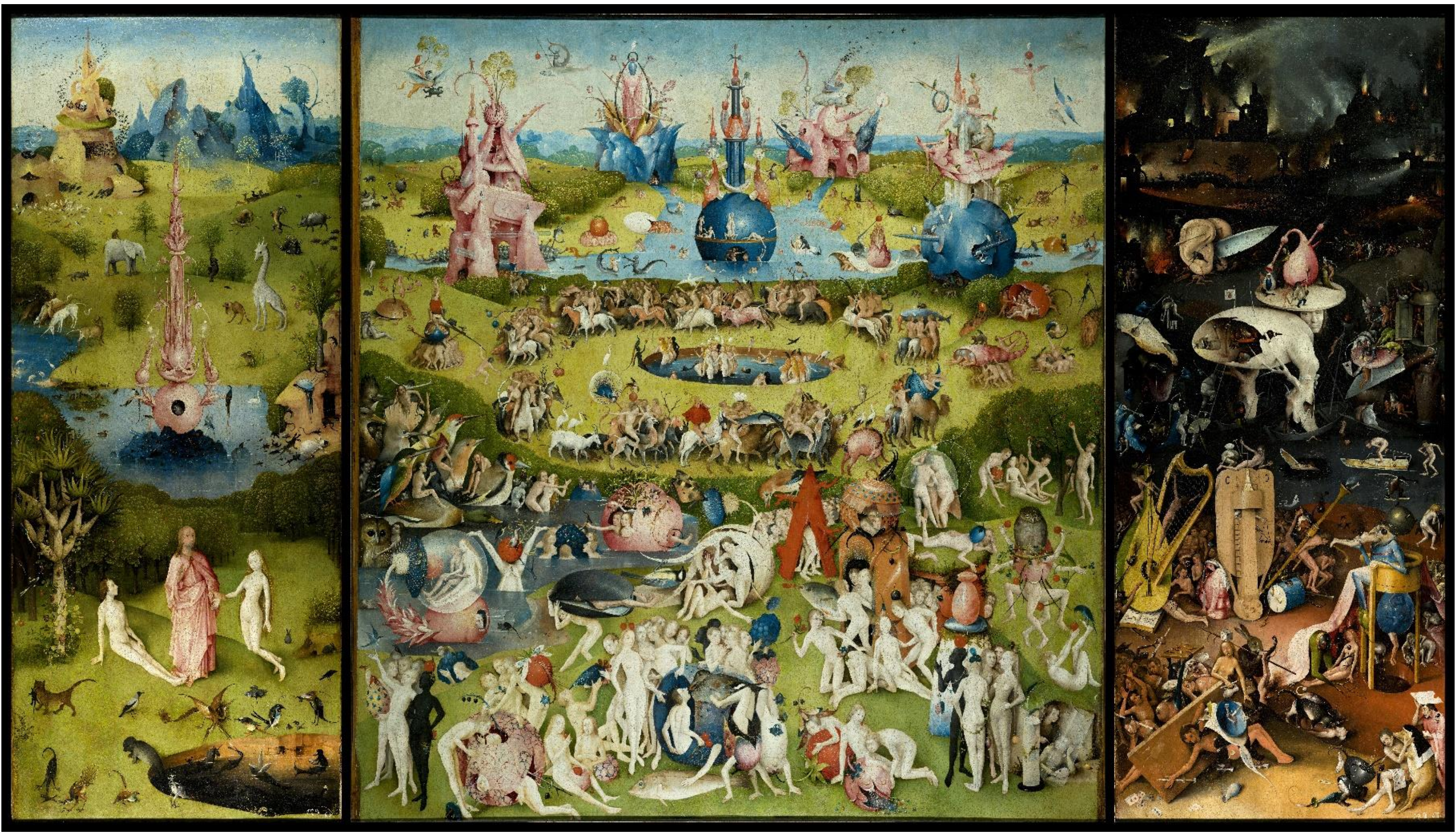
CRITIQUE

An open access journal devoted to the critical study of
philosophy by undergraduates worldwide, published by
Durham University

MMXXII **NEW SERIES** **Summer 2022**

Contents

Articles



The Garden of Earthly Delights, Hieronymus Bosch, 1490-1510

The Garden of Earthly Delights

The Garden of Earthly Delights is a triptych oil painting on oak panel painted by the Early Netherlandish artist Hieronymus Bosch between 1490 and 1510. It is considered one of his most famous works and is housed in the Prado Museum in Madrid, Spain.

The triptych depicts a Garden of Eden-like landscape on the left panel, an allegory of pleasure in the center panel, and a vision of hell on the right panel. The central panel is filled with a surreal and fantastic array of characters, including fantastical creatures, allegorical representations of vice, and scenes of debauchery and sin. The right panel depicts a nightmarish world of torment and suffering, with demons punishing the damned.

The Garden of Earthly Delights is widely regarded as one of the masterpieces of Northern Renaissance art and has been interpreted in many ways over the centuries. Some view it as a moralizing allegory warning against the dangers of worldly pleasures, while others see it as a celebration of human desires and passions. Despite the various interpretations, it remains one of the most enigmatic and intriguing works of art from the period.

Editor’s Introduction

THIS INTRODUCTION is always used to address the state of the journal. They have, recently, lost the optimism which first characterized them with the launch of the new series of *Critique*. This essay is even less optimistic, because this issue of the journal has fallen further away from the dialectical ideal that was first sought for it. By this, of course, I do not mean that the essays published in this issue are of any lower quality. Quite the contrary: although there are only four of them, they are some of the best that have been published by *Critique* in the past few years. Where this issue falls short, however, is in its absences: where are the book reviews? where the discussion pieces? where the back-and-forth between undergraduate students?

This is not a problem unique to this journal. Numerous other undergraduate journals worldwide have, in recent years, made special efforts to solicit discussion pieces focussing on essays in their journals. Here is one serious problem with the way they are doing so: if the piece is rejected, nobody else will publish it. Why is this? Because each journal is soliciting essays that discuss other essays *in that journal*. If that journal will not publish the discussion piece, no other journal will. This is something most academics have experienced first hand in professional journals and, thus, there is a certain risk associated with the authorship of discussion pieces. We have aimed to resolve this issue at *Critique* by soliciting essays discussing undergraduate essays published *anywhere* but, I fear, there is still too great a risk of authoring a discussion piece: at the end of the day, all that means is that there are two rather than one journals that might publish it, and that is still a very high risk to take. Of course, in addition to this, we have the self-obsessed nature of academic research and the structure of undergraduate degree courses – two further issues especially prevalent amongst undergraduates (although the first exists at large amongst professional academics too) that prevent the authorship of discussion pieces. This was the topic of a previous introduction.[1] Consider what I have said here an endnote to that essay.

Likewise, the question of book reviews is still an outstanding one. This is, in part, explained by the structure of university courses, there being an emphasis on the production of research articles and not book reviews. It is also explained in part by the academic climate: book reviews are considered inferior to original research articles; thus, senior academics will have dozens of research articles on their curriculum vitae, but barely a handful of book reviews.

How to get undergraduates writing them? This is, to me, an institutional question, and not one that can be answered by *Critique*. What determines what is written by undergraduates is what will get them a first class degree, and that is out of our control as an undergraduate

[1] See B.V.E. Hyde, (2022), “Editor’s Introduction: Or, A Theory Regarding the Neglect of Reviews and Discussion Pieces, Especially by Undergraduates,” *Critique*, vol. MMXXII, no. 1, pp. i-iii.

journal. As much as I should like universities to have undergraduates review recent books in the field – just as they are these days averting from a historical education and towards one focussed around problematics and the contemporary discussion thereabout – what universities choose to do is not my prerogative.

One of the other shortfalls of this issue about which I am obliged to be frank is its global reach. In previous issues, we have been successful in obtaining submissions from the United Kingdom, United States, Canada, Continental Europe, Asia and even Africa. In this issue, we only received submissions from the UK and US in the first place.

The cardinal aim of future issues of *Critique* will have to be its re-internationalization. This is easier said than done: special efforts have been made in the past with varying success.[2] For one, there is a language barrier, especially when it comes to undergraduate students in Asia, bearing in mind that the essays published in this journal are in English. Of course, nations like Singapore and Hong Kong can be targeted initially, many of their courses being taught in English anyway, but the language barrier still lurks around. Not only does it effect the authorship of essays, but it effects the distribution of the journal's call for papers. We have no idea how well received our communications are by universities but, just like when Western scholars receive an email full of Sinographs, it quickly finds its way into the deleted folder, an English email probably does not last very long in Asian circles either.

I do not want to portray *Critique* as a dying journal. Although short, this issue is not a death knell for it. The new series was an attempt at revival which, as far as I am concerned, was successful. The journal is now in need of another boost, as most journals are within the first couple of years. To this effect, the next issue will be a double issue with both the winter and summer issues combined. I will author a book review for the journal again and make an effort to reinvigorate this section. A new editorial board is in construction too, which I hope to bring a new zest to *Critique*. In the end, a journal is only as successful as its editors and, reaching the close of my tenure with the journal, and busier now than ever, I have run out of time to re-re-invigorate the journal. My task now is to train my successors: the next issue will be testament to the success, or failure, of this final task.

B.V.E. HYDE

[2] See B.V.E. Hyde, (2021), "Editor's Introduction," *Critique*, vol. MMXXI, pp. i-ii.

Searching for Spinoza’s Principle of Sufficient Reason

LEONARDO VILLA-FORTE
Cornell University

SPINOZA’S PRIMARY RATIONALIST COMMITMENT is to the strongest version of the Principle of Sufficient Reason (PSR), or so argues Della Rocca.[1] Della Rocca’s version of the PSR states that everything that exists has an explanation for its existence.[2] On this reading, Spinoza’s philosophical system is driven by an unwavering commitment to uncovering and accepting all of the implications of this principle.

Whether or not Della Rocca’s PSR-first account is an appropriate characterization of Spinoza’s philosophy, this much is clear: Spinoza is committed to some form of the principle, and it plays a central role in Spinoza’s general commitment to universal intelligibility. As Della Rocca puts it, the PSR amounts to Spinoza’s ‘rationalist denial of brute facts’, and that ‘For Spinoza, to be is to be intelligible’.[3]

Textually grounding a PSR-first interpretation immediately runs into problems. Spinoza never explicitly states the principle in those terms; he never flags it as axiomatic or even identifies it as such, unlike Leibniz, for instance. Della Rocca locates it in only two places, arguing that they are the clearest statements of the PSR:[4]

> 1a2: What cannot be conceived through another, must be conceived through itself.[5]
>
> 1p11d2: For each thing there must be assigned a cause, *or* reason, both for its existence and its nonexistence.[6]

For Della Rocca, these two passages don’t merely represent something like the ‘spirit’ of the PSR, but rather are the clearest *expression* of the principle. If this is right, they constitute strong evidence that Spinoza was committed to the PSR as Della Rocca presents it. However, the strength of the ‘no brute facts whatsoever’ PSR places a high burden on these two passages to in fact show that Spinoza held a commitment to the principle as construed.

There is much in Della Rocca’s account that I think is right, but my focus will be on the textual basis for attributing to Spinoza this version of the PSR. I agree with Della Rocca that Spinoza is committed to (a flavor of) the strong PSR – at least to the universal intelligibility of things. My claim is that while Della Rocca is right to identify the PSR in 1p11d2, 1a2 does

[1] Della Rocca 2008
[2] Ibid. 4
[3] Ibid. 9
[4] Ibid. 4-5
[5] Spinoza 1994, 1a2
[6] Ibid. 1p11d2

Critique, vol. MMXXII, no. 2, pp. 1–8.

not state and does not by itself entail Spinoza's PSR as Della Rocca presents it. My goal is to qualify Della Rocca's claim about Spinoza's PSR. I argue that there is better textual evidence for attributing the principle to Spinoza.

A different textual basis for Spinoza's PSR has implications for understanding which flavor of the principle Spinoza was actually committed to (the question of which flavor he should be committed to is not the concern of this paper). First, I will consider the evidence Della Rocca provides for his reading, raise some concerns, and then present a different textual basis for the PSR while offering some reasons why should take, and might even want, the PSR to be in those passages instead.

Why does Della Rocca read 1a2 as a statement of the PSR? He writes, 'For Spinoza, to conceive of a thing is to explain it. Thus, in presupposing in 1ax2 that everything can be conceived through something, Spinoza presupposes that everything is able to be explained, he builds the notion of intelligibility into the heart of his metaphysical system'.[7]

The reading turns on the first claim – Are 'conceive' and 'explain' terms with identical meaning, and as such are interchangeable, for Spinoza? I argue, contra Della Rocca, that they are not. As we will see, the general problem is that the textual evidence for this equivalence seems is unclear, and there are reasons to think the terms come apart in Spinoza's usage.

Della Rocca makes the following three points[8]: 'Spinoza moves freely' between the claims that 'substance is conceived under a certain attribute' and 'substance is explained by that attribute'.[9] Next, 1a5 is taken to suggest that 'conceiving' is the same as 'understanding' a thing, and therefore serves as evidence that 'conceiving a thing' is the same as 'explaining that thing'.[10] Finally, in 2p7s, according to Della Rocca, 'Spinoza says that when we perceive effects through their causes, we are explaining the order of nature. Spinoza sometimes uses "perceives" and "conceives" interchangeably in these contexts (see, e.g., 2-38d)…This shows that Spinoza regards claims about conceiving one thing through another as equivalent to claims about the explanation of one thing by another'.[11] Here, the 'conceive = explain' formula is derived from a 'conceive = perceive' formula, a 'perceive = explain' formula, and the transitivity of identity.

If merely identifying these passages is to serve as evidence for a 'conceive = explain' formula, it must be the case that Spinoza intends phrases that use each to have the same meaning, or uses the terms to make the same point, otherwise this move is question-begging. After all, we would expect these epistemological terms to bear a close relation to each other, and therefore expect to find them being used similarly and in proximity to each other even if we took each term to have a different meaning. It's perfectly possible that the phrases that use each term are true but have different meanings. But, nothing suggests that this is the case, or at least Della Rocca doesn't provide us with a reason to read them as identical.

To start, in 1p10, the use of 'conceived through itself' to discuss attribute could suggest the identity of attribute with substance.[12] In fact, the demonstration that each attribute is 'conceived through itself' might depend on this reading of the attributes. Nevertheless, it's not

[7] Della Rocca 2008, 5

[8] Della Rocca 1996, 3-4; 2008, 4-5

[9] Della Rocca cites 1p10s, 1p14d, 2p5

[10]Spinoza 1994, 1a5: 'Things that have nothing in common with one another also cannot be understood through one another, *or* the concept of the one does not involve the concept of the other.'

[11] Della Rocca 1996, 3

[12] Spinoza 1994, 1d3

clear why the claim that substance has to be explained *by* an attribute is equivalent to the usage in 1p10.[13] Next, I think 1a5 suggests, indirectly, at least that having a concept of something is necessary for understanding that thing, but it's not clear from 1a5 alone whether having a conception of something is sufficient for having a sufficient explanation for it.

Finally, let's consider the 'conceive = perceive' formula. I think it's implausible that Spinoza took conceiving and perceiving to be interchangeable. Moreover, I think this additional formula only increases the implausibility of Della Rocca's reading. The argument begins with the strange claim that 'perceive' is equivalent to 'explain'. While implausibility is normally not a point against attributing a view to Spinoza, we also know Spinoza holds that we perceive everything that goes on in our body (the object of our mind).[14] Even on a charitable reading of what counts as our 'body', it doesn't seem like I have a full-blown explanation, in the sense required for Spinoza's strong PSR, in all cases where I have perception.

Even granting this first step, I'm skeptical to continue by equating 'conceive' with 'perceive'. Della Rocca cites 2p38d. This proposition aims to establish that we only have adequate conceptions of common notions. In the demonstration, Spinoza claims that these can only be conceived adequately, and that the mind also adequately perceives them.[15] A natural reading, to me, of the demonstration and its corollary is that, since there are some things that are in all bodies (such as motion and the ways it changes due to physical laws), each of us can't but perceive them (understood as sense perception), so we will form the same concepts of them. We conceive of A insofar as it is what is common to all, but we perceive A in each particular thing. In other words, conceiving of A as a common notion requires perceiving A in bodies first.

So, pace Della Rocca, it doesn't seem sufficient to identify that Spinoza 'moves freely' between phrases (both 'conceive = explain' and 'conceive = perceive') to establish the 'conceive = explain' formula. These textual troubles leave us with several questions. If we don't adopt the 'conceive = explain' formula, how should we understand Spinoza's use of 'conceiving' and 'perceiving'? Not all of my disagreement about the 'conceive = explain' formula is critical – While I won't provide and defend a fully developed account of any of Spinoza's epistemological language, I want to at least point in the direction of a view by suggesting that 'conceive' and 'perceive' are weaker than 'explain'.

'Perception', as Spinoza uses it, seems close to sense perception as we normally understand it. In particular, Spinoza takes what we perceive to be the 'imaginations of the mind'[16]. And yet, perception understood this way doesn't seem by itself to be able to provide, and certainly not entail, sufficient explanation for the existence of things in the sense required for the PSR.

Next, the relation between 'conceive' and 'explain' might be such that explanation entails conceiving, but conceiving doesn't entail explanation. The mental state of conceiving isn't sufficient by itself to guarantee that the conceiver has a robust explanation in the sense demanded by the PSR. Conceiving, or possessing concepts, might make it possible for us to represent things; i.e., things must fall under certain categories of concepts to be intelligible to us. To say that everything is conceived, as 1a2 does, could mean that everything is such that it can be conceptualized, or categorized into concepts. For instance, that modes are conceived

[13] Ibid. 1p14d, 2p5
[14] Ibid. 2p12
[15] Ibid. 2p38
[16] Ibid. 2p17

under attributes suggests that we must represent modes as instances of certain concepts (Thought and Extension) to understand them.[17] Conceiving, then, serves as a condition for explanation, but it seems to fall short of being sufficient for the explanation to be given. In order to provide a sufficient explanation for a thing's existence, we still must fully understand the causes of the thing we conceive.

'Conceive' has diverse uses, such as conceiving 'under' or 'through', many of which seem weaker than being sufficient for explanation. It would suffice to show cases of conception without sufficient explanation, or without an adequate grasp of the causes of a thing, to bring the two apart. In 1p15s5, Spinoza writes that we conceive of water as divisible and consisting of separate parts 'insofar as it is water, but not insofar as it is corporeal substance'.[18] This example shows that having a certain conception of a thing is not the same as having an explanation *in the sense* Della Rocca requires it to in order to locate the PSR in 1a2. I can conceive of the divisibility of water and that it consists of molecules standing in parthood relations and fail to appreciate how it fits into the picture of the 'corporeal substance'. In other words, I have some conception of water, but I don't possess a full-blown explanation of that water, including its causes, or its relationship to substance, at least not in virtue of (and identical to) my conception.

At this point, denying Della Rocca's reading of 1a2 raises some questions. How should we understand 1a2 instead? Where is Spinoza's PSR? First, we need to understand the relevant kind of explanation (of the existence of a thing) at stake. What counts as a sufficient explanation of a thing for Spinoza? I take 'sufficient reason' for Spinoza to be the set of causes, or the full causal history, that explains why a fact obtains. In 1p11d2, Spinoza identifies 'cause' with 'reason' – The reason for a thing's existence is its causes.[19] Since without one of the causes in the causal history of a contingent thing that thing wouldn't have existed, the 'sufficient reason' for a thing's existence must be its full causal history. Providing the causal history is a sufficient explanation for a thing's existence.

This can also be put equivalently in terms of grounding, or the explanation in virtue of which something obtains. As Della Rocca writes, 'The explanation of a fact is enough – sufficient – to enable one to see why the fact holds'.[20] It is in virtue of the causes, or reasons, of a thing that the thing exists. The causes of a thing, which constitute the sufficient reason, also provide the explanation that grounds a thing's existence.[21] So, explaining a thing's existence requires providing all the causes, or reasons, of a thing. 1a4 confirms this:

> 1a4: The knowledge of an effect depends on, and involves, the knowledge of its cause.[22]

Spinoza holds that everything is caused by God and comes from God, so all explanations ultimately just appeal to God.[23] Since everything is caused, everything is an effect, and so 1a4 can be understood as the claim that the knowledge of anything requires knowledge

[17] Ibid. 1p15d

[18] Ibid. 1p15s5

[19] Ibid.1p11d2

[20] Ibid. 4

[21] Amijee 2020. She writes that, 'because effects are understood through their causes, a causal (and thus conceptual) connection is, for Spinoza, a paradigm case of grounding' (66).

[22] Spinoza 1994, 1a4

[23] Ibid. 1p15.

of its cause. If we take having an explanation to entail, *a fortiori*, having knowledge of what is being explained, 1a4 suggests that explaining anything requires spelling out a thing’s causes.

The use of ‘involves’ might suggest that explanation is just *a priori* entailment, or that the explanation of a thing follows logically from knowledge of its causes. In other words, if we had knowledge of a thing’s full causal history, we would see how its existence follows as a matter of logical necessity. All facts are just logical consequences of God. But, setting aside this complication, we reach another hint that Spinoza intends ‘conceive’ to be weaker than ‘explain’. Notice that Spinoza doesn’t use ‘conceive’ here, but rather ‘knowledge’, or perhaps more accurately, ‘cognition’.[24] Why not? This supports my claim that, for Spinoza, explanation entails conception, but the entailment doesn’t hold in the other direction.

At this point, one can respond by equating, as Della Rocca does, ‘knowledge’ with ‘conceive’ to salvage the ‘conceive = explain’ formula. However, this strategy rules out any usage of a term where we might expect Spinoza to use another one from serving as evidence of different notions intended by these terms. I have a general concern about this. Given the complex relations between all of these related terms, we should at the very least be wary of reducing all of them to mere identity – it doesn’t seem charitable to me to treat them as all interchangeable and with no appreciable difference in meaning or use for Spinoza – do we really want to say, of such a systemic writer that each time we come across one of ‘conceive’, ‘explain’, ‘perceive’, ‘cognize’, ‘understand’, etc, we should treat them as if Spinoza just picked one out of a hat each time he had to use one?

Now, not only does identifying 1a2 as a statement of the PSR lack textual support, but there are also several independent concerns that suggest why wouldn’t want to expend so much effort to squeeze a PSR out of it anyhow. The first is a broader concern that has Spinoza’s geometric method in mind: Spinoza never actually appeals to 1a2.[25] After presenting it, it is never cited or used in any demonstration. To my knowledge, this sets 1a2 apart from all other axioms. By itself, this is a very strange fact.

It would be even stranger, I think, if Spinoza’s master principle was stated in an unused axiom. If Spinoza’s commitment to the PSR is really expressed in 1a2, this might even diminish the importance of the PSR for Spinoza’s system and in relation to all of the important doctrines he derives, with undesirable consequences for Della Rocca’s PSR-driven reading.

Della Rocca might respond by claiming that, in virtue of being so fundamental to Spinoza, Spinoza relies on 1a2 implicitly without ever needing to cite it. In other words, it’s working for Spinoza as a background assumption. I find this to be a strange thing to take Spinoza to have done.

To say that this axiom is ‘in the background’ seems rather vague to me, and, further, uncharitable to such a systematic philosopher as Spinoza, who is using a geometric method no less. Unless implicit uses of the axioms are pervasive, which I don’t take them to be, there are two problems – First, it feels wrong to assume the unique centrality of 1a2, when, for instance, 1a1 seems just as important. Why would Spinoza only treat 1a2 like this, without leaving even a note? Why should this axiom get special treatment, or, what’s the *reason* (since we’re on the subject) this axiom is used without citation, but none of the other axioms are? Second, if it is implied, where is it? I found this question trickier than I initially expected, but the burden of proof lies on those claiming it is used implicitly throughout the *Ethics*. I think it will be difficult to show that any demonstration is glaringly incomplete without an implicit appeal to 1a2.

[24] This was suggested by Prof. Steinberg.

[25] Garber 2015; Vlasits 2021.

Some commentators take 1a2 to be cited implicitly in 1p4.[26]

> 1p4: ‘*Two or more distinct things are distinguished from one another, either by a difference in the attributes of the substances or by a difference in their affections*’.[27]

It’s not clear to me that 1p4 requires an appeal to 1a2, or that Spinoza intended to explicitly reference 1a2. On the one hand, 1a2 *could* possibly have been cited, replacing 1a1, I think, to show that that all that exists is ‘substances and their affections’ by using the second half of 1d3 and of 1d5. The demonstration would otherwise remain unchanged. However, as constituted, the demonstration works without 1a2 because I think the two parts of d3 and d5 are intended to be equivalent statements with the same meaning.

The possibility of citing 1a2 exists only because 1a1 and 1a2 are quite similar. Della Rocca would agree, as he writes of 1a2 that ‘this is equivalent to the claim that everything is either a substance or a mode of a substance’.[28] But, it is eminently clear that the same analysis can be given of 1a1. The appeal to 1a1 in 1p4 shows that Spinoza certainly applies this analysis to 1a1. In trying to appeal to 1a2 in 1p4, we’ve created a problem: 1a1 and 1a2 make the same claim. Do we really want to say that Spinoza’s first two axioms state the same thing in different terms?

I can only conclude that 1a2 is strange enough make me skeptical of identifying it as one of the two clearest statements of the PSR. And, if we claim that the specter of 1a2 is mysteriously present throughout the *Ethics*, our claim is inherently impossible to ground textually, only increasing my skepticism.

In introducing 1a2, I initially claimed that to deny that 1a2 is a statement of the PSR entails that to ‘conceive’ is not the same as ‘explain’. This wasn’t entirely true, because there is a plausible, weaker reading of 1a2 that still blocks Della Rocca’s conclusion. The motivation for this reading comes from the tight connection we’ve identified between 1a1 and 1a2 – They, at the very least, have very similar meaning.

Just as 1a1 begins with a qualifier ‘Whatever is’ on the scope of the claim, it’s plausible that 1a2 should actually be understood is saying something like: [Of whatever is conceived], what cannot be conceived through another, must be conceived through itself.’ On this reading, 1a2 does not entail that everything can be conceived.[29] Further, it doesn’t require a cause, *or* reason, for the nonexistence of anything, so, as we will see, it is a much weaker claim than the statement of the PSR in 1p11d2.

At this point, there’s one question left. If not in 1a2, where is the PSR? I think we can find Spinoza’s PSR stated much more explicitly and directly in other passages, including 1p11d2. This is the second passage Della Rocca uses as his main textual support for his PSR-first reading of Spinoza. It reads: ‘For each thing there must be assigned a cause, *or* reason, both for its existence and its nonexistence’.[30] Since everything that exists is either substance or mode, and all things (*res*) have explanations, everything that exists (and everything that doesn’t exist) has an explanation.[31] For reasons above, I take this to be a stronger statement of the PSR.

[26] Vlasits (2021) offers an argument in favor of this.

[27] Spinoza 1994, 1p4

[28] Della Rocca 2008, 70

[29] Schneider (2014) reads 1a2 this way in denying that the PSR is in any of the Part I axioms, and suggests it is weak enough to allow that something could come into existence without a cause, or reason.

[30] Spinoza 1994, 1p11d2

[31] Ibid. 1a1, 1p15, 1p11d2

Spinoza doesn’t cite any axioms or propositions established before 1p11, so it seems he just assumes the PSR. This passage also helps us potentially find a statement of the PSR even earlier, in 1a3, which reads: ‘From a given determinate cause the effect follows necessarily; and conversely, if there is no determinate cause, it is impossible for an effect to follow’.[32] This axiom basically claims that nothing can be uncaused – everything has a cause. Taken together with 1p11d2, 1a3 holds that everything has a reason – this closely resembles the PSR.

Finally, I take the PSR to be present in 1p8s2, which claims ‘that there must be, for each existing thing, a certain cause on account of which it exists’.[33] While this passage doesn’t explicitly require causes, or reasons, for nonexistence, it is almost identical to the PSR in 1p11d2. Since Spinoza subsequently takes this principle to imply that there must be a reason for why a group of individuals is *not* a different number of individuals, I think he takes it to be, in effect, just as strong as 1p11d2.[34] Again, Spinoza offers no justification for the principle.

The implication of my textual evidence for the PSR is that it is not located in the axioms but is something like a self-evident truth[35]. One advantage of my reading of Spinoza’s PSR is that in the best-case scenario for 1a2’s PSR, it only emerges as the downstream effect of reading the axioms through the lens of conclusions established in subsequent propositions – So, instead of starting with the PSR, this reading makes the PSR almost a derivative result of the propositions that follow – this doesn’t seem to be the way Della Rocca would like to characterize Spinoza’s fundamental, prior commitment to the PSR.

At this point, one could object: This textual evidence, unlike 1a2, is unable to ground Della Rocca’s claim that Spinoza’s central, most fundamental commitment is to the PSR. It is true that this reading denies that the PSR is in any of the axioms alone of Part I, but I don’t think this is much of a problem. We don’t lose anything by finding the PSR outside of the axioms given that we’re already attributing to Spinoza a principle he never explicitly states or labels. There are two additional points to make in response to this concern. First, the strangeness of 1a2 suggests that the grass isn’t always greener on the other side (in axiom-land, at least). Second, the PSR, as I see it, is a self-evident truth for Spinoza. What exactly this means and the relation between axioms and self-evident truths is unclear.[36] However, it would be hard to deny that Spinoza, in calling the PSR self-evident, acknowledges at least some fundamentality, close to that of an axiom, to the PSR.

Finally, the high burden that we identified for what counts as a sufficient explanation and my reading of 1a2 raises a concern about a frequent pattern of inference that Della Rocca employs throughout, namely his ‘explicability arguments’.[37] The argument appeals to our inability to explain something, like the non-identity of two relations or of substances with different attributes, and concludes by rejecting the existence of what we can’t explain. However, a sufficient explanation for a thing’s existence, as we saw, must spell out the causal history of a thing. Further, we have only confused, inadequate knowledge of most things.[38] If there is something we can’t explain that looks to us like a brute fact, what justifies our inference

[32] Ibid. 1a3

[33] Ibid. 1p8s2

[34] Ibid. 1p8s2

[35] Ibid. 1p11d2*:* ‘These things are evident through themselves…’

[36] Schneider (2014) reaches a similar conclusion and offers interesting answers to both of these problems.

[37] See Della Rocca 2008; 2010

[38] Spinoza 1994, 2p13s

to the denial of the existence of that thing?[39] Della Rocca's argument pattern seems to conflate between the psychological claim that I can grasp explanations for everything and the metaphysical claim that there are explanations for everything.

One final objection to this endeavor would claim that Della Rocca might appreciate and even support my claims about textually grounding Spinoza's PSR in other places, so it might be reasonable to ask how much we really disagree, or if any of this matters in the end. After all, I think Della Rocca is correct when he identifies 1p11d2 as a statement of the PSR. However, the disagreement is substantive and relevant in virtue of the different textual bases we offer for the PSR. Further, analyzing whether the PSR is in 1a2 comes with the added benefit of improving our understanding of 1a2 and the relevant terms. After all, 1a2 is a very unclear and puzzling axiom.

References

Amijee, Fatema. "Principle of Sufficient Reason," *The Routledge Handbook of Metaphysical Grounding*, ed. Michael Raven, 63-75. New York: Routledge, 2020.

Della Rocca, Michael. *Representation and the Mind-Body Identity in Spinoza*. Oxford University Press, 1996.

Della Rocca, Michael. *Spinoza*. Routledge, 2008.

Della Rocca, Michael. 'PSR', *Philosophers' Imprint* 10, no. 7 (2010):1-13.

Garber, Daniel. 'Superheroes in the History of Philosophy: Spinoza, Super-Rationalist', *Journal of the History of Philosophy* 53, no. 3 (2015):507-521.

Newlands, Samuel. *Reconceiving Spinoza*. Oxford University Press, 2018.

Schneider, Daniel. 'Spinoza's PSR as a Principle of Clear and Distinct Representation', *Pacific Philosophical Quarterly* 95, no. 1 (2014):109-129.

Spinoza, Baruch. *A Spinoza Reader: The Ethics and Other Works*. Edited by Edwin Curley. Princeton University Press, 1994.

Vlasits, Justin. 'Everything is conceivable: a note on an unused axiom in Spinoza's *Ethics*', *British Journal for the History of Philosophy* (2021):1-12.

[39] Samuel Newlands (2018, 31) writes that, 'The PSR is not indexed to human intellectual capacity'. Newlands rejects the PSR-driven reading on this concern alone, claiming it would be uncharitable to Spinoza since this argument from explanation is fallacious. But, if it turns out the reason for something being inconceivable to us is that it is inconceivable in principle, this inference from explanation might be justified even if what we can conceive and everything conceivable aren't coextensive.

Two Stories in One – Thought Experiments, Selfhood, and What Williams Missed Out

MARTA BAX
Cambridge University

THOUGHT EXPERIMENTS feature heavily in the literature concerning continuity of self. Reflecting on the functional role they fulfil for such discussions, Bernard Williams (and some like-minded contemporaries)[1] has concluded thought experiments reveal an arbitrariness to the debate on whether the self is continuous with psychological or bodily facts. The unsettling entailment which follows is that 'selfhood' is a conceptual matter dependent on our contingent intuitions. Therefore, one is free to accept either theory on the basis of whichever intuition they deem strongest. I reject this conclusion. Thought experiments, particularly those employed by Williams, establish two *non-arbitrary* accounts of selfhood that do not rely on intuitions, but on the state of affairs obtaining in the actual world. Only one of these continuity accounts will correspond to the state of affairs as they are. Therefore, only one account will emerge victorious.

The argument employed against Williams is strongly reminiscent of a discussion in philosophy of language concerning the role intuitions play in confirming theories of reference. In presenting this critique of Williams', I draw parallels to Max Deutsche's rejection of intuition-based claims proposed by the 'experimental philosophers'.[2] The conclusions he reaches in 'Experimental Philosophy and the Theory of Reference'[3] will help to elucidate and inform those we ought to reach in (thought) experimental philosophy and the (continuity) theory of self.

§1

Thought experiments offer-up answers to a very specific *about*-question[4] in philosophy of mind centring around the continuity conditions of the self. These discussions envision the self as a mobile entity capable of being transferred from one body to another, or *psychologically continuous*; *or* one which is fundamentally continuous with a physical body, and therefore not transferable between two or more bodies, describing *bodily continuity*.[5] Thought experiments appeal to intuitions as evidence in favour of any given theory, and so have been termed

[1] See Beck, 2014.
[2] Deutsche, 2009, p. 445.
[3] Ibid.
[4] Question about the persistence conditions of the self as opposed to the identity conditions.
[5] See Shoemaker, 1984, p. 109; Olson, 1997.

Critique, vol. MMXXII, no. 2, pp. 9–19.

'intuition pumps'.[6] If the pump succeeds in establishing an intuition compatible with psychological persistence conditions of the self, then one may consider an account of selfhood which allows for such a transfer to take place. If instead we are led to think of the self as bodily continuous, then its description must follow in such a way for it to be impossible for the self to persist after the death of said body.

The pre-Williams set-ups favour intuitions supporting the psychologically-continuous notion of self. These follow specifically from the *body-swap* thought experiment proposed originally by John Locke.[7] Any formulation of the body-swap, Locke's in particular, asks whether 'the self' — composed of psychological properties such as memories, thoughts, and feelings — can be preserved during the transfer from one body to another. Updated modern — typically physicalist— accounts instead make use of *brain-transplant* experiments (cerebrum transplants for some)[8] to reach similar conclusions. These examples go one step further by locating the self in the brain. The thought is that where the brain moves, psychologically continuous thoughts, memories and feelings, follow, as does the self by extension. The difference is that they envisage the brain (or part of), which supports distinctive psychology, being transferred, rather than simply transferring the psychology itself. Insofar as the possibility of transfer is plausible, body and brain-swap experiments in this vein have been read as *confirmers* for a psychological continuity theory of self.

Williams offers his own:

> *Body Exchange Machine*
> 'Suppose it were possible to extract information from a man's brain and store it in a device while his brain was repaired, or even renewed, the information then being replaced'.[9] Now 'suppose that there were some process to which two persons, *A* and *B*, could be subjected'[10] in which case 'information extracted into such devices from *A's* and *B's brains* (is) replaced in the other brain'[11] so that 'there is a certain human body which is such that when previously we were confronted with it, we were confronted with person *A*, certain utterances of it were expressive of past memories of *A*; but now, after the process is completed, utterances coming from this body are expressive of what seem to be just those memories which previously we identified as memories of the past experiences of B; and conversely with the other'.[12] Faced with the prospect of being tortured post-process,[13] *A* selects *B* to receive the cash prize, on the assumption that the body swap has taken place and his own self will inhabit the *B*-body, and vice-versa for *B*.[14]

[6] Dennett, 2013.
[7] Feser, 2007, p. 66-68.
[8] Shoemaker & Swinburne, 2003, p. 9.
[9] Williams, 1970, p. 162.
[10] Williams, 1970, p. 161.
[11] Williams, 1970, p. 161-2.
[12] Williams, 1970, p. 161.
[13] Williams, 1970, p. 163.
[14] Ibid.

Intuitions in this case work in a way which favours the psychologically-continuous narrative over the bodily-continuous. However, the same thought experiment is re-formulated in order to appeal to completely opposite intuitions:

> *Forgotten Torture*
> ‘Someone in whose power I am tells me that I am going to be tortured tomorrow... He adds that when the time comes, I shall not remember being told that this was going to happen to me, since shortly before the torture something else will be done to me which will make me forget the announcement... He then adds that my forgetting the announcement will be only part of a larger process: when the moment of torture comes, I shall not remember any of the things I am now in a position to remember... He now further adds that at the moment of torture I shall not only not remember the things I am now in a position to remember, but will have a different set of impressions of my past ... (he adds) lastly that the impressions of my past with which I shall be equipped on the eve of torture will exactly fit the past of another person now living, and that indeed I shall acquire these impressions by (for instance) information now in his brain being copied into mine.’[15]

It seems now that *A*’s, *B’s* and our own intuitions on whether a body-swap has occurred ought to look wildly different. On the prospect of being tortured post-memory-transplant: ‘Fear, surely, would still be the proper reaction: and not because one did not know what was going to happen, but because in one vital respect at least one did know what was going to happen-torture, which one can indeed expect to happen to oneself, and to be preceded by certain mental derangements as well.’[16] For Williams, the non-mobile-self intuition established by *Forgotten Torture* is naturally reached by most competent English speakers. If so, this scenario completely undermines the functional role that has historically been attributed to body-swap thought experiments.

Here’s the crux: There must be *some* difference between *Body-Exchange Machine*, and *Forgotten Torture*, given the two stories pull us in opposite intuitive directions. Examining the cases, we find changes in the language employed such as the tense shift from third person to first, or the delay in mentioning the ‘other man’ in *Forgotten Torture*.

However, such differences are non-substantively trivial in the sense that preferring one re-telling over the other is an arbitrary matter – we have no reason to prefer the first-person re-telling over the third other than intuition. If the deciding factor warranting preference of one pump over the other comes down to intuition alone, then there is no reason to favour *Body-Exchange Machine* over *Forgotten Torture* other than with an appeal to whichever intuition comes out strongest. But that in itself will come down to a matter of personal preference for the inquirer. Williams opts for the *Forgotten Torture* re-telling, and thus the immobile-self intuitions it gives rise to. But he concedes that this choice is an arbitrary one made on the basis of intuitions which he deems strongest *from his personal perspective.*

An attempt to point out non-trivial differences between the experiments to identify a non-intuition-based reason for choosing one pump over the other is therefore (a) impossible and (b) irrelevant to the thought-experiment project. It is impossible because the reason for our

[15] Williams, 1970, p. 168.
[16] Williams, 1970, p. 168-9.

intuition-changes between scenarios are explained via trivialities in language, method or formulation. It is irrelevant because *the fact that trivialities in formulating thought experiments make all the difference to our final intuitions on whether the self is mobile is exactly what thought experiments are used to show.* For any psychologically continuous account of the self, an 'equivalent' thought experiment can be formulated as a falsifier. This is what 'Forgotten Torture' does. The exact opposite is achieved for any bodily continuity theory of self, as shown by 'Body Exchange Machine'. In other words, the function of thought experiments is to provide indirect support to a particular theory 'by showing that its opposition is in trouble'.[17] They 'illustrate that we don't *have* to see things in the way a particular theory says we do'.[18]

This is interesting because it points to a potential recipe for turning mind-swap cases into body-swap cases. In fact, Williams' steps to focus on the *A* self through the use of first person and the focus on one self only could feature in this hypothetical manual. The approach is to an extent reminiscent of Linda Zagzebski's strategy in 'The Inescapability of Gettier cases'.[19] Given the right language used, one can fathom turning *any* fabricated instance of knowledge into a justified true belief. This is another case where direction of intuition is heavily affected by the wording of a thought experiment alone.

Let's take stock. *Body Exchange Machine* and *Forgotten Torture* are versions of the same body-swap thought experiment, albeit described in different terms which agitate two inconsistent accounts of selfhood. *Body Exchange Machine* describes the world as one where the self is psychologically continuous. *Forgotten Torture* reflects the world where the self is bodily continuous. These differences are non-conclusive, given their triviality. Hence, whichever of the two states-of-affairs correspond to the actual world is determined by the contingent, personal preference of the inquirer. Therefore, whether the self is mobile is a conceptual matter dependent on our contingent intuitions. Thought experiments appeal to these different intuitions, and hence output conflicting conclusions about the self's nature. Therefore, there is no good reason to accept one version of selfhood over another, unless by appeal to intuition. This does not lend conclusive support towards a theory of self as compatible with psychological continuity as opposed to bodily continuity, or vice versa. Enquirers are therefore free to adopt either theory about the self.

§2

A similar argument features in philosophy of language, specifically theory of reference. This time, variations of intuition are applied to Kripke's case against a descriptivist theory of meaning for proper names. According to the descriptivist, an ordinary proper name refers to an object identified by the definite descriptions a speaker *S* associates with its name. If you know me as 'the writer of this paper', then referring to my name will mean you refer to the writer of this paper. Kripke employs a thought-experiment which elicits intuitions opposing this particular theory of reference.[20] We imagine a case where unbeknownst to us, it was not Gödel, but Schmidt, who proved the incompleteness of mathematics. We then imagine a speaker who associates the name 'Gödel' with the single description 'the prover of incompleteness'. Opposing the conclusion derived from a descriptivist theory of meaning, Kripke's case would

[17] Beck, 2014, p. 194.
[18] Ibid.
[19] Zagzebski, 1994.
[20] Kripke, 1972, p. 83-84.

have us conclude that when this speaker uses 'Gödel', they are really referring to Gödel —of whom they do not realise the definite description does not apply—not Schmidt. If this is right, then descriptivism fails. Kripke's thought experiment *falsifies* descriptivism in this way.

The experimental philosophers[21] resist Kripke by appeal to the arbitrariness of intuitions conjured up by thought experiments. Empirical data shows that on the *same* retelling of the Gödel-Schmidt story, East Asian intuitions about reference support the conclusion that when the speaker uses 'Gödel', they are really referring to Schmidt.[22] Kripke's case would not be a falsifying example against descriptivism for speakers who share such intuitions. This case lends even greater evidence towards the arbitrariness of selecting a theory by intuition given that here the two opposing intuitions arise from the same word-for-word retelling of a story — no need for trivial differences in language — informed by cultural relativity. Since there is no reason to adopt an ontological view of reference informed by Western intuitions over East-Asian ones —other than by appeal to unjust bias—Kripke's thought experiment fails to falsify or confirm descriptivism as our best theory of reference. Instead, one is yet again free to accept either theory on the basis of whichever intuition they deem strongest.

The general position which applying to both Kripke and Body-Swap proponents is this: that their cases motivate people to accept specific versions of events is not evidence for accepting that these events do obtain, because significantly similar cases (Williams) or the same case (Kripke) will motivate the opposite conclusion for different people. And a premised argument capable of entailing two opposing conclusions (that the self is both psychologically and bodily continuous) does not provide us with any interesting philosophical information other than perhaps the note that thought experiments as a method compromise the very theories that rely on them to work.

§3

This next section develops the position against the arbitrariness claim. I provide an explanation to the following insights: First, the state of affairs described by Williams' thought experiments are not-so-trivially different from each other. Secondly, whichever account is favourable will be determined by the state of the world-as-it-is *independently from* our intuitions, rather than reliant on them. In the Williams' case, only one of *Body Exchange Machine* and *Forgotten Torture* will correspond to the state of affairs as they are. Therefore, only one conceptualisation of the self as psychologically continuous vs bodily continuous is correct.

It may first be useful to make clear the distinction meant by '*stories'* and '*facts'* they latch onto. In Kripke's case this is straightforward: his *story* is the case of Schmidt and the meddling Gödel, and this thought experiment is mobilised as evidence for the *fact* that 'Gödel' refers to 'Gödel'. Stories highlight the intuitions one might have about independently obtaining affairs in the actual world. The issue is that this story elicits intuitions that support two different states of affairs: in the first, Gödel refers to Gödel, in the second, Gödel refers to Schmidt. And given a name cannot refer to multiple people, we have a problem.

The application of 'stories' and 'facts' to the selfhood case manifests slightly differently. Williams' insight is that although *Body Exchange Machine* and *Forgotten Torture* are literally distinct in terms of language used; conceptually, the way the text represents the world is the same: memories are transferred from one brain to the other. There is one story at

[21] Deutsche, 2009, p. 1.
[22] Mallon et al., 2009.

play, albeit trivially described in two versions. The language / story distinction is similar to the way 'Clark Kent' and 'Superman' can refer to the same person. The names are clearly different in a literal sense, but the question is whether they pick out the same thing/process in the world. And it seems like they do, just as the re-telling of *Body State Transfer* as *Forgotten Torture* so does. Then, again the problem is that the same story supports intuitions pointing to two separate and conflicting ways such a process could describe the world to be: one where the self is psychologically continuous vs bodily continuous.

Kripke's account with this in mind so closely resembles Williams' that examining how to deal with the charge against him may be instructive to navigating Williams. Deutsch's reply to the experimental philosophers on behalf of Kripke is to deny that there is only one story at work. There is a specific detail that goes unconsidered in the original case making all the difference. The question (Q) 'Who is John referring to?' can be interpreted in two *non-trivially distinct* ways:

(Q1) To whom does *the name, 'Gödel', refer* when John uses it?[23] *Semantic reference*
(Q2) To whom does John *intend* to refer when he uses 'Gödel'? *Speaker reference*

Deutsche's bet is that the pooled East Asian speakers who answering accordance with descriptivism misinterpreted 'Who is John referring to?' to mean (Q2), where Kripke, and non-descriptivists, seek an answer to (Q1). If this is true, then East Asian speakers, and indeed anyone who interprets (Q) as (Q2), have *misunderstood* the thought experiment as intended by Kripke. Instead of responding to the story about semantic reference, these speakers respond to one concerning speaker reference. Once clarified, their referential intuitions will not differ significantly from Westerners, and so similarly converge on the conclusions reached by the correct theory of reference which operates in the actual world.[24]

It is important to pause here to emphasise that 'East Asian intuitions' and 'Western intuitions' are purely contingent labels with regards to the kind of intuitions about reference people might experience with regards to this case. In fact, I find that my intuitions, when asked 'Who is John referring to' side more with the speaker reference interpretation, and I am not an East Asian speaker. What the argument rests on is an agreement that the meanings of (Q1) and (Q2) may come apart, and so generate seemingly conflicting answers. Perhaps the distinction then may still appear unintelligible to some, *and they may be correct.* It may be the case that some people think that 'Gödel' only ever refers to the person the speaker intends to refer to. Specifying this fact, however, still provides an explanation as to why people's intuitions seemed to differ so wildly in the first place.

So, it is mistaken to think that people's intuitions differ by default, and that there is no further evidence to discriminate between positions. It is rather a product of ambiguous wording that some justifiably misunderstand Kripke to be telling a different story than the one he actually is, one that naturally reflects different facts concerning reference.

[23] Deutsche, 2009, p. 454.

[24] Further question on why theories established as true fall in line w Western intuitions – i.e. why is a theory of reference built around the (Q2) interpretation as opposed to the (Q1) one?

§4

How does this apply to the body-swap thought experiments? The list of differences in language between the two versions of the body swap story as pointed out by Williams —tense shift; use of first person — is not exhaustive. Specifically, there is a 'trivial' difference in formulating *Brain State Transfer* and *Forgotten Torture* responsible for the unchecked emergence of the two *significantly different stories* (S1) and (S2).

The different intuitions which consequently arise concern questions about how the psychology of a person is preserved and whether such preservation is possible in the absence of a causal chain.

I propose that the language presented in *Body State Transfer* reflects (S1) where the memories and other transferable phenomenology of persons *A* and *B* are *originals* which have survived the brain-state-transfer process. It is not controversial to read Williams this way, given he purposefully introduces a specific mechanism – the very brain-state-transfer-device – which is described as working in exactly this way:

> suppose it were possible to extract information from a man's brain and store it in a device while his brain was repaired, or even renewed, the information then being replaced.[25]

Body Exchange Machine so implies that mental information —the mobile candidate for our concept of selfhood— is preserved in an *uninterrupted chain of existence* from its residence in *A's* physical body, to *B's* new body. This concept of transfer via preservation comes out when contrasted with the idea of transfer by re-introduction of non-original memories. Consider an analogy:

If I promise to buy my sister a specific toy in its packaging she has seen in a shop and named 'Pat' on the Sunday, but on the Monday walk in to find that the exact object she had pointed to had been sold already, I do not give her 'Pat' when I gift her the next best option, an exact copy of the toy, the next day. I give her a non-original copy. Whether she believes this to be Pat or not, and whether the intruder has most properties Pat has, makes no difference to whether this object is identical with Pat – it simply is not. The object identical with Pat is the one being carried home by the early-bird stranger who took the packaging before I had the chance.

A transfer story involving 'stored' memories means the conditions of identity imposed on the *A*-self during transplantation require spatio-temporal continuity as well as descriptive continuity. It is not sufficient that the memories of *A* transplanted into body *B* are in content the same as those held previously by body A eg. both body *A* and *B* have a memory of playing in the park. Such memories must be physically identical in that if body *B* now has the memory of playing in the park, body *A* lacks this memory. When conceiving of transplant experiments between two bodies, it is the case that there is only one set of original memories that is moved so that its introduction in one body means its elimination in the other.

The incorporation of *Brain State Transfer Device* suggests that there is something about the original memories — ones caused by a real interactive experience with the environment — which critically contributes to selfhood. Given these memories may plausibly transcend from one body to another, selfhood has psychological continuity as a feature. If *A's* original

[25] Williams, 1970, p. 163.

memories are transported to body *B*, *A's* self too undergoes transportation. This is the story of selfhood implied by (S1) *Body Exchange Machine*.

Forgotten Torture importantly lacks the *Brain State Transfer Device* in its narrative. The transfer of self in *Forgotten Torture* is described like so:

> I shall not only not remember the things I am now in a position to remember, but will have a different set of impressions of my past ... (he adds) lastly that the impressions of my past with which I shall be equipped on the eve of torture will exactly fit the past of another person now living, and that indeed I shall acquire these impressions by (for instance) information now in his brain being copied into mine.[26]

Note the emphasis is on having sets of impressions which have been *copied* onto a brain. In light of the conversation following intuitions elicited by *Body State Transfer*, it should be clear how *Forgotten Torture* blurs the lines on exactly which type of memory-transfer has taken place. 'Copy' is semantically closely related to non-originality. The doll I bring home is an exact *copy* of the one my sister wished for. The 'copying' of memory impressions may easily lead one to conceive of *non*-original memories making up the phenomenology of the post-transplant *B*-bodied person.

Induced memories perfectly resemble their original counterparts in content. These are non-delusional memory-like experiences not causally connected to remembered events via experience, but by some encoded memories reflecting an experience that was not had.[27] The memories transferred in *Forgotten Torture* are induced-memories, not originals, in virtue of their being *copied*. If this is the case, it is crucial to the discussion given we may raise doubts as to whether a transfer of self via preservation of original memories vs one where original memories are destroyed and consequently copied point trivially to the same process, or state of affairs in the world. And if this is the case, then the case reflected by *Forgotten Torture* cannot just be a simple re-telling of (S1).

This is crucial because it explains why it is that intuitions about the continuity of self change when confronted with this case. Given original memories are not preserved, it is natural that someone who takes originality of memory as essential to selfhood concludes no transfer of self has taken place. Person *A* in question is continuous with their body, because no matter what induced memories are copied into his brain— be them derived from some other person or completely new — they lack the originality 'stored' or preserved by transfer. In fact, given his own original memories have been wiped, there is nothing *but* bodily continuity that can guarantee *A*'s preservation post-process.

If this is the case, then the ambiguity 'copying' elicits in a reader is comparable to the ambiguity 'refer' would in the Kripke case. Just as the clarification on whether Kripke means semantic or speaker reference in his case makes all the difference to the conflict of intuitions which arises, so does the specification on whether by 'copying' one means preserved transfer, or deletion and recreation. And if this is the case, it is no longer surprising at all that different intuitions arise from the two cases: one is entitled to feel two different ways about two very different processes. You *can* think that a self can only be transferred is preservation of original memories is involved, and so your intuitions about *Body State Transfer* and *Forgotten Torture* will necessarily differ.

[26] Williams, 1970, p. 168.

[27] Schechtman, 2011.

§5

We may wonder whether it's possible to re-write *Forgotten Torture* story in a way that avoids the issue for Williams. Deletion and re-creation of memories by being copied into a brain may be considered equivalent to being downloaded onto a computer and then re-uploaded back into a brain. This may entail some kind of preservation in the (S1) sense. But there are two ways of countering this approach: (1) Deny the equivalence. Downloading and uploading via a computer that can store a memory does not involve deletion. In fact, this case is more a re-telling of *Brain Exchange Machine*, with the computer replacing the *Brain Sate Transfer Device*, than it is of *Forgotten Torture.* Given deleting and preserving are substantially different notions, the original point made —that *Brain State Transfer* and *Forgotten Torture* reflect two different stories, not versions of one — stands. (2) Embrace the equivalence. Downloading and uploading is just a version of deletion and re-creation, one that fails to preserve the original memory in an appropriate manner. Turning the memory into code, and then inputting code into a new brain, *is* the same as copying an originally deleted memory, turning it into an induced memory.

Additionally, we cannot maintain that psychological states are deleted and re-created whilst holding the psychological view because there is no guarantee of *continuity* if, in the moment of deletion and before re-creation the self (if psychological) ceases to exist. The only thing maintaining continuity would be the continued existence of a body and so bodily continuity is implied.[28] Someone like Derek Parfit is happy to accept a story of selfhood where it is the case that a person may come in and out of existence.[29] However, this reply is disallowed in the context of this particular debate because the appeal is to a type of view of the self which does not take continuity as necessary. But the debate we are concentrating on – one concerning psychological vs bodily *continuity* – already assumes this as a primary requirement.

In fact, it is important to note that continuous accounts of selfhood such as the ones considered hitherto are not the only accounts of selfhood available. Eliminativist accounts (see Hume,[30] Metzinger,[31] Johnston)[32] are unaffected by this line of reasoning. They are anti-realist in the sense that there is no continuity discussion to be had because there is no self which continuity can attach to. But clearly, this paper is restricted to philosophical claims which pre-suppose the existence of the self as a continuous entity.

The copying process is therefore significantly different to the transfer process described in *Body-State-Transfer*, and so reflects a story (S2) where self-transfers are carried out in this non-psychology-preserving manner.

It is therefore no surprise that the same intuitions which lead proponents of the psychologically-continuous view to accept *Body-State Transfer* lead them to accept *Forgotten Torture.* One can maintain that self-transfer has taken place in the former and not the latter by virtue of the substantial difference that proper preservation of real memories only occurs in *Body State Transfer*. The intuitions arising from both cases — in the first there is transfer, in

[28] In the case of the computer, one would be committed to saying that downloading preserved memories onto the computer means that you are the computer, a view that is not impossible to hold, but might be intuitively resisted by some

[29] Parfit, 1984, p. 119.

[30] Hume, 1964.

[31] Dainton, 2012, p. 162.

[32] Dainton, 2012, p. 171.

the second there isn't — do not conflict, and both support the fact that the self is psychologically continuous.

Philosophers who favour the story told by *Forgotten Torture* similarly do not fall into a contradiction if intuition pushes them to accept a transfer has taken place in *Body State Exchange.* They may maintain that *if* the *Body State Transfer Device's* existence were plausible, then transfer of self is open for consideration. But such a device would never (a) exist, or (b) work in a way that is desirable for proper memory transfer to take place. Given that when we talk of transfer, we actually mean copying of induced memories, it is safer to accept an account of selfhood based on bodily continuity.

From here it does not follow that the facts described by *Body State Exchange Machine* and *Forgotten Torture* both obtain in the actual world. (S1) assumes a world-view where the self if psychologically continuous, and (S2) assumes a world where it is bodily continuous. The aim of the essay was to show that these two positions are not just arbitrarily based on intuition. In order to determine whether by 'reference' one means semantic reference (Q1) or speaker reference (Q2), specify as to whether one is identifying an object as *intended by the speaker or not*. In order to determine whether by 'continuous self' one means with-psychology or with-the-body, specify as to whether we mean transfer of actual memories (S1) or copies of induced memories (S2) in thought experiments.

The next natural step in theorising on selfhood, then, is to ask whether one can conceive of the brain-state-transfer device and its function as a real life possibility. The notion of copying memories seems pretty feasible already – we transfer various software from one body to the other every day in the world of computing. Perhaps it is this feasibility which leads Williams to side with the bodily-continuous account of selfhood towards the end of his paper.[33] It is not that his justification is stronger albeit arbitrary intuitions he has which pull him towards accepting *Forgotten Torture* over *Body State Transfer*. It is that he believes the state-of-affairs described by *Forgotten Torture* and obtaining in (S2) reflect those obtaining in the actual world. This belief is further justified by the reflection that a transfer of psychology is metaphysically possible only via the copying of memories, rather than a physical transfer of the original memories themselves.

In summary, Williams' contribution to the debate on selfhood is an argument mandating suspicion of using thought experiments to establish any proper evidence in favour of the psychological or bodily continuity positions concerning selfhood. At best, intuition pumps work to discredit these positions. In this essay I have shown (a) that this conclusion is wrong, and (b) that value of thought experiments in this debate stretches beyond falsification. Thought experiments do not falsify respective continuity accounts of selfhood because the problematic conflicting intuitions which arise from them turn out not to be conflicting at all. Moreover, pumps importantly disambiguate crucial terms used in the selfhood literature. *Body Exchange Machine* and *Forgotten Torture* particularly highlight that the notion of 'transfer' reflects two significantly different states of affairs in the world. Determining which of the two obtains, they inform us, is our next step.

[33] Williams, page 180.

REFERENCES

Beck, Simon (2014). Transplant Thought-Experiments: Two costly mistakes in discounting them, *South African Journal of Philosophy,* 33(2):189-199.

Dainton, Barry (2012). Self-hood and the Flow of Experience, *Grazer Philosophische Studien,* 84 (1):161-200.

Dennett, Daniel C. (2013). *Intuition pumps and other tools for thinking*. Allen Lane.

Deutsch, Max (2009). Experimental philosophy and the theory of reference, *Mind and Language*, 24(4): 445-466.

Feser, Edward (2007). *Locke.* Oneworld Publications.

Hume, David (1964). *A Treatise of Human Nature.* Aldine Press.

Kripke, Saul (1972). Naming and Necessity, *Tijdschrift Voor Filosofie*, 45(4):665-666.

Mallon, R., MacHery, E., Nichols, S., & Stich, S. (2009). Against arguments from reference, *Philosophy and Phenomenological Research*, 79(2), 332-356.

Shoemaker, Sydney (1984). *Personal Identity*. Blackwell.

Shoemaker, Sydney; Swinburne, R. (2003), *Personal Identity*. Blackwell.

Schechtman, Marya (2011), 'Psychological continuity theories In: Personal identity' In: *Routledge Encyclopedia of Philosophy,* Taylor and Francis.

Olson, Erik T. (1997). Was I Ever a Fetus? *Philosophy and Phenomenological Research,* 57(1), 95–110.

Williams, Bernard (1970). The Self and the Future. *The Philosophical Review*, 79(2), 161–180.

Zagzebski, Linda (1994). The Inescapability of Gettier Problems. *The Philosophical Quarterly*, 44(174), 65–73.

Is the Denial of the Existence of an Ultimately True Theory Self-Contradictory?

GIACOMO BARTOLESCHI
Oxford University

> Whatever is dependently coarisen,
> That is explained to be emptiness
> That, being a dependent designation
> Is itself the middle way.[1]
> —MMK XXIV.18

According to the Madhyamaka doctrine of emptiness all things are devoid of intrinsic nature (svabhāva). If “ultimate truth” is analysed as correspondence with how a thing is intrinsically, then no theory can be ultimately true in at least this sense. Garfield and Priest 2003 argue that this leads to the contradiction “the ultimate truth is that there is no ultimate truth”.[2] Their conclusion however will be argued not to follow because the denial of the existence of an ultimately true theory is not ultimately true but it is nevertheless really true. Even if this may not be an exegetically accurate reading of Nāgārjuna,[3] this essay will argue that the denial, the doctrine of emptiness, and theories in general can express *true relations* even if they do not express ultimate (that is, intrinsic) facts about relata. Relational facts will be argued to be conventional but really true, so no contradiction arises. The structure that this essay will follow to argue that the denial of the existence of an ultimately true theory is *not* self-contradictory is the following: section 1 will analyse the sense of “ultimate truth” relevant to the denial, section 2 will argue that the denial is not ultimately true in that sense, section 3 will then argue that the denial can consistently be really true and section 4 will finally argue that the doctrine of emptiness too (besides the denial) can be really true without being an ultimately true theory.

1. What is the “ultimate truth” relevant to the denial?

To clarify what sense of “ultimate truth” should be understood as relevant to the denial, the doctrine of emptiness and the doctrine of the two truths will be briefly outlined. According to

[1] Mūlamadhyamakakārikā as translated by Garfield, 1995 p. 69
[2] Garfield and Priest, 2003 p. 13, quoting Siderits, 1989
[3] Compare for instance MMK XV rejection of extrinsic nature (parabhāva) in Garfield, 1995. pp. 220-224

Critique, vol. MMXXII, no. 2, pp. 21–29.

the doctrine of emptiness all things we grasp (such as objects, concepts, sensations, and even ourselves) co-dependently arise, so nothing can be rationally established to have intrinsic nature. Being empty, śūnyatā, is thus being devoid of svabhāva, a term which may be translated as inherent existence, substance, essence, or intrinsic nature.[4]

If "there is nothing non-empty" (Vigrahavyāvartanī v.23)[5] it follows that no theory, including the doctrine of emptiness, can express how anything is *intrinsically*. For instance, if the pot I see doesn't actually have an intrinsic nature, a theory such as "this pot is blue" cannot correspond to what really is the case about the pot in itself. The terms used in theories, such as "pot" and "blue", are conventional designations which cannot be established to correspond to any intrinsically existent object or property. If we understand "ultimate truth" as correspondence to the intrinsic facts about things, the doctrine of emptiness denies the existence of an ultimately true theory in at least in this sense.

MMK XXIV.8 says that the teaching of the Buddha should be understood in terms of two satya, a Sanskrit term meaning truth and/or reality: paramārtha-satya, reality and/or "truth in the highest sense", and saṃvṛti-satya, which means either veiled or conventional truth and/or reality.[6] The conventional (not necessarily veiled) sense of saṃvṛti, according to Nāgārjuna's commentator Candrakīrti, can itself be understood in three ways: (a) nominal or by linguistic convention, (b) relative, and (c) empirical.[7] Garfield and Priest, amongst others, generally translate paramārtha-satya as ultimate truth,[8] but this essay will argue that this is not the same sense of "ultimate" denied by emptiness, that is correspondence with intrinsic facts. In this essay paramārtha-satya will be understood *not* as the ultimate truth which is denied but rather as what is fundamentally (not necessarily intrinsically) really the case independently of linguistic conventions.

This essay will *not* argue that nothing is really the case independently of linguistic convention. This view is arguably hardly intelligible, and it would be quite difficult to avoid the contradiction that at least something would really be the case, namely that nothing really is the case. Moreover, as Garfield and Priest themselves observe, emptiness for Nāgārjuna "is emphatically not nonexistence but, rather, interdependent existence".[9] This essay will thus argue that the above understanding of the doctrine of emptiness is consistent with holding that some relational facts really are the case. Since, as it will be analysed in detail later, the identification of relata is at least dependent on a subject, relations are arguably conventional in the sense of (b) relative or (c) empirical, but *not* necessarily conventional in the sense of "veiled" or (a) nominal. Theories can still express relations that are true independently of linguistic convention. The denial will thus be argued to be really true but not ultimately true, so it is not self- contradictory.

2. THE DENIAL IS NOT ULTIMATELY TRUE

In order to explicitly deny the existence of an ultimately true theory we must either utter terms such as "theory" or, at least, say "no" or silently shake our head, and the meaning of these

[4] Garfield, 2002 p.24
[5] As translated by Westerhoff, 2010 p.28
[6] Garfield and Priest, 2003. p.5
[7] Ibid.
[8] Ibid.
[9] Ibid, p.6

expressions is set by convention. Even if the denial's formulation depends on linguistic convention, however, it may still express something that is really the case independently of linguistic convention. Compare two familiar examples: on the one hand the denial of the existence of a meter longer than 4 feet clearly says nothing about convention-independent facts. On the other hand, however, denying that pots fall upwards seems to express a fact about reality (or at least about our subjective experience) which doesn't seem to depend on the convention of what is "up". If we change linguistic convention the truth value of the old statement may change, but we still experience pots to fall in the same direction. The denial of the existence of an ultimately true theory, roughly similarly to this second example, will be argued to point at a truth independent of linguistic convention, but not an ultimate truth.

The denial may be argued to be contradictory if it expressed an ultimate truth about theories in themselves, but this is not the case. According to the doctrine of emptiness there is no ultimate truth about theories because "theories" do not exist intrinsically. We usually grasp theories as an object of thought or perception, but according to Nāgārjuna this doesn't imply that theories exist intrinsically. In MMK XXII.9-10 he claims "whatever grasping there is / Does not exist through essence [...] Grasping and grasper / together are empty in every respect".[10] According to the doctrine of emptiness whatever is grasped (theories in this case) is interdependent on the subject who grasps it, and thus they are both empty. If we accept Nāgārjuna's doctrine of emptiness theories dependently co-arise with the subject, and are thus devoid of intrinsic existence. If there is no intrinsically existent theory the denial cannot correspond to what is ultimately the case about theories.

It may be objected that instead of expressing an ultimate truth about theories the denial expresses an ultimate truth about what really exists, but this is also not tenable. The non-existence of theories is not an intrinsic fact about what exists, it at least depends on what "theories" are. Since according to the doctrine of emptiness there is no such thing as what "theories" ultimately are, their absence or difference from what really exists cannot be an ultimate fact. Therefore the denial cannot be ultimately true about reality either.

3. The denial can consistently be really true without being ultimately true

Even if the denial is not expressing any ultimate truth about either reality or theories in themselves, it will be argued to express a true relation involving empty theories. Nāgārjuna's arguments on causal relation and differences seem to give some credibility to this position, but even if the this weren't an exegetically accurate reading of Nāgārjuna it will be argued to be a consistent position to hold nevertheless. Firstly in Vigrahavyāvartanī.22 Nāgārjuna says that empty things are causally efficacious: things "which are empty of substance because they are dependently originated, perform in their respective ways",[11] and in MMK VIII.6 too he asserts that effects are real, even if interdependent on subject, causes and conditions: "If there are no effects, liberation and / Paths to higher realms will not exist. / So all of activity / Would be without purpose".[12] While the nature of causal relations is beyond the scope of this essay these passages suggest that Nāgārjuna is not denying that empty relata can be connected by real relations.

[10] Garfield, 1995

[11] Westerhoff, 2010 p.27

[12] Garfield, 1995 p.24

The doctrine of emptiness doesn't straightforwardly deny the reality of relations, it rather asserts that relations cannot be established to be grounded in intrinsic facts about relata. When in MMK XIV.5 Nāgārjuna says: "a different thing depends on a different thing for its difference",[13] he is not claiming that the differences we perceive in conventional (in the sense of empirical) reality are unreal. Since Nāgārjuna rejects nihilism, denying difference would imply that empirical reality is uniform, which is a difficult position to hold, and not his view. What can be consistently claimed is rather that if emptiness is true things cannot be established as intrinsically different, but they are really different relatively to an observer and each other.

Holding that theories only express true relations is consistent regardless of whether or not there are also ineffable intrinsic facts that ground those relations. On the one hand it may really be the case that the true relational facts expressed by theories are not grounded in any intrinsic facts about either relata. While a satisfactory defence of this first position is beyond the scope of this essay, it is generally accepted to be the case for constant motion. Special Relativity is generally interpreted as suggesting that there are no *intrinsic* facts about constant motion, but there are real facts about *relative* motions.[14] Whether a pot is moving relatively to me is a fact and it doesn't seem to be grounded in any intrinsic fact about either my motion or the motion of the pot in itself. There are no intrinsic (that is, ultimate) facts about the motion of a single thing in itself (like the pot), but there are objective facts about relative motions. If this interpretation of Special relativity is correct, conventional truths in the sense of (b) expressing relative facts can thus be objectively true.

We can, moreover, conventionally decide to describe all motions relatively to a conventional reference frame, in which case there would also be convention-dependent objective truths about the motion of the pot, and still no ultimate (intrinsic) truth about it. While it may be objected that in the future the above interpretation of Special Relativity will turn out to be inadequate, or that it picks out the only special class of relational facts which are not grounded in intrinsic facts, it at least shows that some theories express empirically adequate relations without also expressing any intrinsic (ultimate) fact. The consistency of the denial, moreover, will be argued to follow even if there are ultimate facts but theories cannot express them.

If there is some ineffable ultimate truth the denial of the existence of an ultimately true theory can consistently express a true relation between theories and ultimate truth. The denial of ultimately true theories, in fact, doesn't require that there are no *ineffable* ultimate truths. Although Nāgārjuna is sometimes interpreted as claiming that there is no ultimate reality,[15] his philosophical arguments on emptiness can at best show that ultimate reality cannot be conceptually grasped and rationally established by a subject. Ultimate reality may for instance be unveiled (but neither grasped nor established by reason), by relinquishing all views and the cessation of conceptual grasping. The denial may thus be interpreted as claiming that the ultimate is not expressed in theories because theories are views grasped by thought.

Śāntideva, in Bodhicaryavatara 10 for instance claims: "The ultimate is not grasped as an object of thought; Thought is explained to be merely conventional".[16] This claim may be objected to seemingly transcend the same limits of expressibility it draws, contradictorily expressing something true about the ultimate, namely that it is not grasped as an object of

[13] Ibid.
[14] See for instance Steane, 2012.
[15] Compare Garfield and Priest, 2003.
[16] As translated by Vaidya, 1960

thought. No contradiction arises, however, since the claim expresses no ultimate truth about reality. Not being graspable by thought is not an *intrinsic* property of the ultimate, but rather a *relation* between the ultimate and thought. Since according to the doctrine of emptiness thought doesn't ultimately exist "the ultimate is not grasped as an object of thought"[17] is conventionally true.

Moreover, regardless of whether there are or not ultimate facts it is consistent to hold that the denial doesn't express any intrinsic fact but it is really true. Consider for instance the theory "blue *looks* different from red". This theory doesn't express any intrinsic fact about either blue or red. The nature of the visual experience that "blue" designates may in fact vary substantially for each person, even if we generally all agree that it looks different from "red". "Blue" and "red" only conventionally designate two experiences without describing their look or expressing their intrinsic nature, they are empty designations. The theory "blue looks different from red" is nevertheless really true, it expresses an intersubjectively evident phenomenological fact that holds independently of linguistic convention. Therefore even if as claimed by Śañtideva thought is conventional, it can still conceptually grasp a true *relation* (like difference) at the empirical level without grasping the intrinsic (ultimate) facts that may or may not underlie it. It is thus consistent to hold (as argued in previous paragraphs) that the denial expresses nothing intrinsic about either theories or ultimate reality, but still expresses true relations between these empty relata. The denial is in this sense really true but not ultimately true, so no contradiction arises.

4. The doctrine of emptiness is not an ultimately true theory

If the doctrine of emptiness were an ultimately true theory the above arguments would not avoid the contradiction. This because the doctrine of emptiness is one of the premises used by this essay to argue that we can consistently claim that there are no ultimately true theories, so if it were an ultimately true theory it would undermine the conclusion. It will however be argued that the doctrine of emptiness is true but not ultimately true, so no contradiction arises in this way either.

Emptiness ends, but doesn't answer, the search of ultimate truth

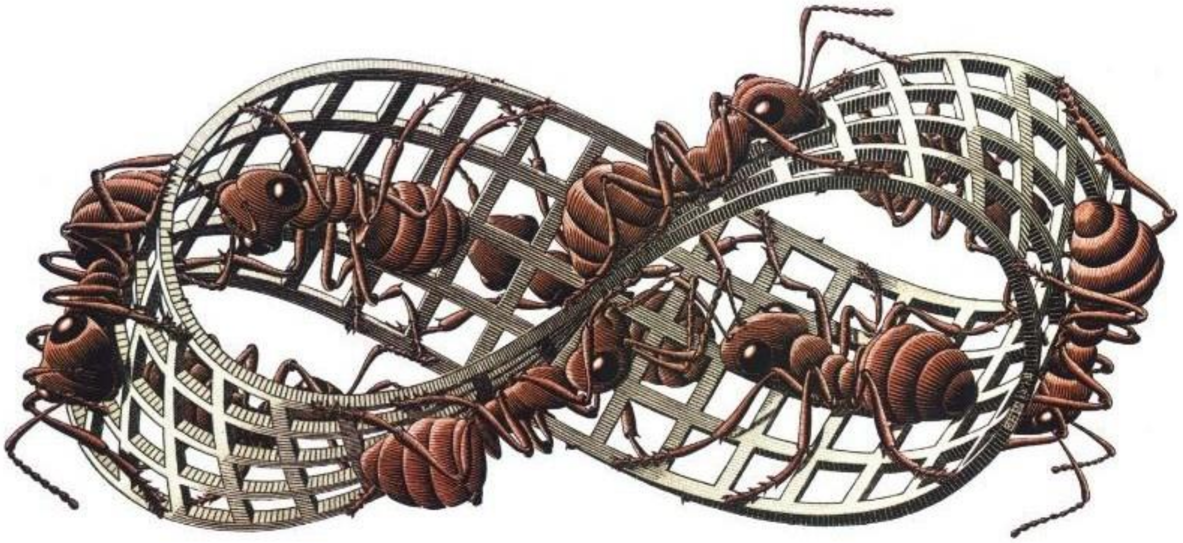
Garfield and Priest 2003 claim that emptiness is the ultimate truth about things. They argue: "for something to be an ultimate truth is for it to be the way a thing is found to be at the end of an analysis of its nature" and "the result of this ultimate analysis is the discovery that all things are empty and that they can be no other way. This, hence, is an ultimate truth about them".[18] If we follow their definition of ultimate truth, however, understanding emptiness doesn't provide an answer to our search for ultimate truth, but rather reveals that the way they formulated our search of ultimate truth was mistaken. If we understand ultimate truth as "the way a thing is found to be" it is not surprising that a contradiction may arise because this definition begs the question against emptiness, which denies that there is a way *a thing* is in itself independently of convention. To avoid the contradiction we can simply understand truth as what really is the case (regardless of "things"), as argued in the previous sections. Therefore even if, as Garfield

[17] Vaidya, 1960

[18] Garfield and Priest, 2003 p.12

and Priest 2003 claim, a contradiction followed from their definition, it would only imply that their specific definition of ultimate truth leads to a contradiction and should thus be rejected in favour of different understanding of truth, and thus the denial would be consistent.

Emptiness, moreover, can really be the case without being the ultimate truth as Garfield and Priest define it. Understanding emptiness, in fact, both reveals that the way we defined ultimate truth was mistaken, and *also* provides an answer to the analysis of what is really the case (regardless of "things" in themselves). To illustrate how this may be, imagine an ant walking on a Möbius strip willing to measure its length (what really is the case). Being used to measuring everyday objects, she believes that length is only found by searching for the two furthest ends of an object and measuring their distance (which represents the search of ultimate truth about things as Garfield and Priest understand it). The ant's usual strategy is doomed to fail when applied to measuring the length of the Möbius strip. Similarly to how the ant can't possibly measure the distance between the ends of the strip, according to emptiness we cannot reach the intrinsic nature of a thing because, in both cases, there is no such thing.



Escher, M., 1963. Möbius Strip II (Red Ants). Available at: <https://www.researchgate.net/figure/MC-Escher-Mbius-Strip-II-Red-Ants-1963-woodcut_fig16_326834755> [Accessed 14 April 2021].

This however doesn't imply that nothing really is the case, that is, that the Möbius strip has no length. Imagine that the ant leaves a mark on one point of the strip and by passing it again it realizes that it has walked the complete length of the strip. The rational ant stops, because it understands that searching further for the ends of the strip is pointless (similarly to understanding emptiness, which reveals there is no ultimate truth as Garfield and Priest define it). The ant's analysis has thus come to an end: she has *not* found what her usual measuring method was looking for (the ends of the strip), but she has nevertheless also found what really is the case about the length of the strip. The ant's assertion "the strip has no ends, and it is really x cm long" may seem contradictory to other ants which still mistakenly assume that having a length requires having ends, but it is in fact consistent and really true. Similarly, our analysis into the nature of something can end without reaching the *intrinsic* nature of that thing. Understanding the mistake of our initial assumption (there are things, and they are intrinsically in some way) it is consistent to say that we fail to find the ultimate truth but we understand what really is the case, and no contradiction arises.

EMPTINESS IS NOT THE INTRINSIC NATURE OF THINGS

While the example above is useful to get a rough intuition of how we might consistently reach an answer to what really is the case and also understand we asked the wrong question, the limitation of the above example is that by using an object to represent reality it suggests that length (what really is the case) is an intrinsic property of the Möbius strip, that is, it is ultimately true. Emptiness, however, is not ultimately true because it is not the intrinsic nature of reality or things. Emptiness may be expressed as “all things are interdependent”, “each thing is devoid of intrinsic nature”, or “each thing is empty” and in none of these senses it can be ultimately true.

Firstly if “empty” is understood in terms of interdependence “all things are empty” doesn’t imply that any one thing is *intrinsically* (that is ultimately) found to be empty. Interdependence involves at least two things, so it doesn’t imply that a single object is intrinsically empty. As discussed earlier using the example of the Theory of Relativity, it isn’t clearly the case that all relations must be grounded in intrinsic facts, so it is consistent to hold that the interdependence of things is not their intrinsic nature. It thus doesn’t follow that the ultimate truth about any individual thing is emptiness in this sense.

Secondly, if we understand empty as devoid of intrinsic nature no contradiction follows either. “Things are devoid of intrinsic nature” is at best conventionally true, because ultimately there is no such thing as an intrinsically existent “thing” or “intrinsic nature”; they are both empty designations. As discuss earlier for the denial, this doesn’t imply that “things are devoid of intrinsic nature” is false, it rather expresses something that is really the case at the conventional level.

Thirdly it cannot be meaningfully claimed that each thing is found to be intrinsically empty. Nothing else than emptiness itself can be established to have emptiness as its intrinsic nature. Suppose for instance that a subject analyses the nature of a pot, eventually finds nothing substantial,[19] and thus calls her final finding “emptiness”. First of all according to the doctrine of emptiness, what she calls “emptiness” is empty, since it designates the subject-dependent conceptual reification and grasping of whatever she found, not *that* in itself. Moreover even if we supposed that “emptiness” ultimately corresponded to that emptiness itself, what would “pot” ultimately correspond to? For “the pot is empty” to be an ultimately true theory, *both* “empty” and “pot” must correspond to something she found at the end of her analysis. However by hypothesis she hasn’t found anything *else* than the nothing she called emptiness. The “pot” only exists at the conventional level, so “the pot is empty” is conventionally true.

The only way in which “the pot is empty” may be claimed to be ultimately true in the above example is if both “empty” and “pot” designated *that same emptiness* she found at the end of her analysis. In this case however “the pot is empty” would just mean “emptiness is emptiness”, which isn’t a theory expressing anything about ultimate reality, it is true regardless of what “emptiness” refers to. Therefore even supposing that emptiness is the ultimate reality, no theory can express anything about it except that it is itself. Since saying that emptiness is

[19] A similar example is provided by Garfield and Priest, 2003 p.12

ultimately emptiness expresses nothing about the way reality is, no theory, including the doctrine of emptiness, can express an ultimately truth.

Conclusion

The denial of the existence of an ultimately true theory is therefore *not* self-contradictory. According to the doctrine of emptiness the denial cannot express any ultimate truth about either theories or reality in themselves, because on the one hand "theories" do not have an intrinsic nature, and on the other hand being inexpressible by theories is not an intrinsic property of ultimate reality. The denial is nevertheless really true because it expresses a true relation between empty relata. This is consistent both if there really are no intrinsic facts about either relata (as suggested with the example of relative motions) and also if there are intrinsic facts about relata, but they cannot be grasped or expressed by theories (as suggested with the example of different colours). The denial moreover is consistent because the doctrine of emptiness as well, which is one of the premises this essay used to argue for the denial's consistency, can be really true without being an ultimately true theory. Understanding emptiness ends our search of ultimate truth without answering it, and it shows that it is really true that theories can never grasp any ultimate fact. Since both the denial and the doctrine of emptiness can be true without being ultimately true theories, the denial of the existence of an ultimately true theory is consistent.

References

Primary Sources

Garfield, J. and Priest, G., 2003. *Nāgārjuna and the Limits of Thought*. *Philosophy East and West*, 53(1), pp.1-21.

Garfield, J., 1995. *The fundamental wisdom of the middle way*. New York: Oxford University Press.

Garfield, J., 2002. *Empty Words, Buddhist Philosophy and Cross-Cultural Interpretation*. New York: Oxford University Press.

Siderits, M., 1989. *Thinking on empty: Madhyamaka anti-realism and canons of rationality*. Rationality in Question.

Steane, A., 2012. *Relativity made relatively easy*. Oxford, United Kingdom: Oxford University Press.

Vaidya, P., 1960. *Bodhicaryavatara of Santideva*. Darbhanga: Mithila Institute of Post-Graduate Studies and Research in Sanskrit Learning.

Westerhoff, J., 2010. *The dispeller of disputes*. New York: Oxford University Press.

Secondary Sources

Deguchi, Y., Garfield, J. and Priest, G., 2008. The Way of the Dialetheist: Contradictions in Buddhism. *Philosophy East and West*, 58(3), pp. 395-402.

Diamond, C., 1988. Throwing Away the Ladder. *Philosophy*, 63(243), pp.5-27.

Garfield, J. and Westerhoff, J., 2015. *Madhyamaka and Yogācāra*. New York: Oxford University Press.

Garfield, J., 2015. *Engaging Buddhism*. New York: Oxford University Press.

Katsura, S. and Siderits, M., 2013. *Nāgārjuna's Middle Way*. Boston: Wisdom Publications.

Siderits, M., 2018. *Buddhism as Philosophy*. New York: Routledge.

Tanaka, K., 2013. Contradictions in Dōgen. *Philosophy East and West*, 63(3), pp.322-334.

Williams, P., 2009. *Mahāyāna Buddhism*. London: Routledge.

Liberating Free Will from the Shackles of Moral Responsibility

ROBBIE BELL
Edinburgh University

THIS ARTICLE DEFENDS A POSITION in the free will debate that is occupied, as far as I have encountered, by Bruce Waller alone. The account revolves around two central claims:

1. **Humans are not morally responsible** – Humans cannot come to *deserve* punishment or reward through their actions in such a way that these responses produce any non-instrumental good (i.e. setting aside deterrence, rehabilitation etc.)

2. **Humans have free will** – The agential capacities possessed by humans are sufficient to count them as having free will (but all the same insufficient to ground moral responsibility)

Gregg Caruso dubs this position 'Reverse Semi-Compatibilism' (RSC).[1] In some ways, RSC is deeply radical. It generates steep moral demands and requires a potentially drastic revision of the conceptual relationship between free will and moral responsibility. On the other hand, this position retains and attempts to resolve conflicting intuitions from both major philosophical camps: incompatibilist concerns over how it could be fair to punish somebody for something that they were 'inevitably' going to do, and compatibilist claims that we have all the agency we could want. If successful this position might be considered a sort of middle-way, and therefore not so radical after all.

RSC, as outlined by Waller, involves at least one further core claim that is beyond the scope of this article:

3. **Moral responsibility practices are not instrumentally justifiable** – Existing systemic practices that rely on moral responsibility would need significant revision if evaluated on purely instrumental grounds

It remains an empirical question how much, and whether, current practices would need changing on instrumental grounds.[2]

[1] Whilst Bruce Waller does not use this term, I do so because it clearly contextualises the view

[2] For discussions of this see Waller, 2011, pp. 133-152; Caruso, 2017; Martinez, 2017; Pereboom, 2006, pp. 158-186

Critique, vol. MMXXII, no. 2, pp. 31–52.

In “Accounts of Moral Responsibility” I establish the form of moral responsibility being denied. I outline arguments for why other accounts of responsibility should not be considered ‘moral responsibility’ at all. This includes responsibility as something akin to ‘attributability’, PF Strawson’s ‘reactive attitudes’ account, and ‘forward-looking’ moral responsibility. In “Moral Responsibility Skepticism” I outline and defend Galen Strawson’s ‘Basic Argument’ against moral responsibility. It argues that humans must be *causa sui*, i.e. self-caused, in order to be morally responsible. I survey an array of compatibilist and libertarian arguments and argue that they all fail to overcome ‘the basic argument’. In “Free Will Preservationism”, I claim that humans still have free will despite lacking moral responsibility. After establishing two criteria of success for the preservation of ‘free will’, I argue that the agential capacities we possess are sufficient to meet them both. I therefore claim that we *can* liberate free will from the conceptual shackles of moral responsibility. I conclude that Bruce Waller’s reverse semi-compatibilism is a viable and attractive position within the free will debate.

Accounts of Moral Responsibility

The sort of moral responsibility that this account denies is the sort that ‘provides the moral justification for singling an individual out for condemnation or commendation, praise or blame, reward or punishment.’[3] When somebody who fits the criteria for moral responsibility acts wrongly, they *deserve* blame and punishment in some non-instrumental sense. Even in the absence of practical benefits, proportional punishment produces some intrinsic good, or at least is considered ‘just’. Something akin to this conception is shared, with some defending it and others denying it, by Aristotle, Spinoza, Nietzsche, McKenna, Galen Strawson, Clarke, Van Inwagen, Wolf, Kane, Mele, Sommers, Pereboom, Nagel, Hobart and many more. It is this sense of moral responsibility that drives Kant’s intuition that, even in the breakup of an isolated island society, a murderer should not be left to live harmlessly in isolation; they *must* be executed.[4] It appears to underly the folk intuitions of most non-philosophers and has a central place in the morality of Abrahamic religions and the criminal justice systems of many secular societies. This view is clearly common, but this dissertation later defends a revisionary account of free will. It is therefore important to show that the above account of moral responsibility is the only valid one, and so by refuting it we rule out moral responsibility entirely.

A key group of alternatives to this desert-based view holds moral responsibility to be something akin to ‘appraisability’, ‘answerability’, or ‘attributability’.[5] Waller’s main criticism of these accounts is that they substitute the hard question of whether somebody is morally *responsible* for an easier one that leaves our main concerns unaddressed. Take Shoemaker’s ‘attributability’, and the ‘aretaic appraisals’ that come with it.[6] Whether we can judge somebody to be ‘selfish’ or ‘cowardly’ in light of an action is blatantly morally relevant. However, it seems strange to suggest that when we speak of moral responsibility, all we are talking about is the correct application of moral judgments of character. Waller uses Gary Watson’s example of the murderer Robert Harris to emphasise this.[7] Is Harris morally

[3] Waller, 2011, p. 2
[4] 1790/1996, p. 158
[5] Zimmerman, 1988; Smith, 2007; Scanlon, 1998
[6] Shoemaker, 2011
[7] Watson, 2018, p. 131

responsible for the murders he committed, given his brutal character was formed through an eye-wateringly violent and abusive childhood? This is a difficult and morally important question but establishing an aretaic appraisal gets us no closer to answering it. We already know that Harris is cruel. The question is whether he's *responsible* for that cruelty. Under the 'attributability' account the general question "can humans be morally responsible?" devolves into something far more trivial like "can humans have morally relevant character traits?". This removes any significance from whether humans as a whole are morally responsible. The justification for preserving free will that I explore later is that our will is as free as we should want it to be. This *prima facie* appears central to what we might mean by 'free will'. By contrast, these accounts bypass entirely 'the basic just deserts question that has driven centuries of concern and debate over moral responsibility'.[8] For a more detailed discussion of these types of accounts, and the particularities of why each should be rejected, see "Redefining Moral Responsibility" in *The Stubborn System of Moral Responsibility*.[9]

P.F. Strawson puts forward another influential account of moral responsibility. According to Strawson, 'our natural human commitment to ordinary inter-personal attitudes... is part of the general framework of human life, not something that can come up for review'.[10] Moral responsibility is inextricably tied with our reactive attitudes, for example anger and guilt. These are so essential for our existence as social animals that they cannot be denied, and therefore moral responsibility can't be denied either. One type of response to Strawson attacks the claim that these reactive emotions are essential to humanity by showing the possibility and preferability of alternative reactive attitude. One such alternative is Derk Pereboom's 'moral sadness'.[11] Similarly, the article "Resentment and Reality" explores the Buddhist alternative reactive emotion, 'lovingkindness'.[12] This indicates that a large population has historically lived in a way that Strawson precludes as impossible or undesirable. Another type of response denies that the reactive emotions are incompatible with moral responsibility skepticism in the first place.[13] For a more general response to Strawson's position, see *Against Moral Responsibility*.[14] Ultimately Strawson's position does not seem a viable account of moral responsibility. Even if Strawson's arguments can justify *treating people* as morally responsible, it feels wrong to say that what *actually makes somebody* morally responsible is other people's emotional responses, rather than some property of the actor themselves. Just like the above accounts, this misses the crucial question of whether we *deserve* those reactions.

The other major alternatives to desert are forward-looking accounts of moral responsibility.[15] The general idea is that somebody is morally responsible when punishing or rewarding them for their actions is likely to have a beneficial effect, for some accounts on their character specifically. Smart describes two cases where a schoolkid, Tommy, did not complete his work. In one case it was because he was stupid and in the other because he was lazy. According to Smart, 'Stupid Tommy' should not be considered responsible, as punishment would not force the stupid out of him. 'Lazy Tommy' *is* morally responsible because the teacher can help produce better future behaviour by punishing him. Importantly, he notes that

[8] Waller, 2014, p. 10
[9] Waller, 2014, pp. 9-38
[10] 1962/2003, p. 83
[11] 2006, pp. 89-99
[12] Goodman, 2002
[13] Nichols, 2007
[14] Waller, 2011, pp. 203-209
[15] Hume, 1748/2000; Schlick, 1930/1939; Smart, 1961

'By this [ascription of moral responsibility] he will not necessarily mean to deny that Tommy's behaviour was the outcome of heredity and environment'.[16] Moral responsibility is justified by the effectiveness of the punishment, rather than punishment being justified by somebody's responsibility. The first way to interpret this is that, given these pragmatic considerations, the person *genuinely is* morally responsible. However, Waller points out that 'For one thing, [this] will imply that some profoundly committed human criminals and terrorists—who are unchanged by even severe punishment—will not be counted as morally responsible, but dogs, rats, and pigeons certainly will be'.[17] This is a strange definition. It ascribes moral responsibility based on properties that don't seem remotely related to 'responsibility'. A morally responsible criminal could free themselves from moral responsibility by training themselves to be numb to punishment. Similarly, an entirely innocent man would be considered morally responsible for something they were falsely accused of if the punishment would produce pragmatic results, such as genuine character growth. This seems an utter butchering of the term 'moral responsibility', describing something more like 'susceptibility to punishment and reward'. See Waller for a more detailed rejection of this redefinition.[18]

A second, more palatable interpretation of forward-looking accounts is a form of pragmatically justified illusionism, most notably outlined by Smilansky.[19] That is, we retain the standard, desert-based definition of moral responsibility and act like we have it when we don't. This is justified on intuitive pragmatic bases, such as that without punishment far more people would act immorally, and immense harm would be caused. For example, Daniel Dennett says 'praise and blame, reward and punishment, are as necessary for civilization as food and water are for life. That doesn't make punishing the guilty "intrinsically good" in the extreme sense that Kant defended, but it does make it a very important good, a practically necessary good'.[20] Even if true, this actually involves *accepting* my definition of moral responsibility, rather than establishing an alternative. So, if successful, RSC *would* show that there is no moral responsibility (we just might have to pretend otherwise). In reality though, removing desert would put immense pressure on forward-looking justifications, so even this weaker, illusory interpretation may still come out false. If people are not actually morally responsible, then the suffering inflicted on them to sustain a system of punishment should be weighed just as strongly as anybody else's. As of 2016 over 10.35 million people were incarcerated globally, with the largest share of this in the United States of America.[21] Add to this all of those suffering psychologically and physically from a culture of blame, as explored in Marshall Rosenberg's *Nonviolent Communication*. This is an enormous moral cost for something like stability and moral incentivisation, which we might be able to produce in other ways. There is a clear moral duty to examine whether the suffering justified by (illusory) moral responsibility *is* in fact necessary, explore alternative systems, and minimise suffering of 'the guilty' wherever possible.[22]

Desert-based moral responsibility is the only account that captures something vital to our intuitions and the justifications for our practices. Even if somebody were to argue for the

[16] Smart, 1961, p. 302
[17] Waller, 2014, p. 13
[18] Waller, 2014, pp. 10-14
[19] Smilansky, 2000
[20] Dennett, 2021
[21] World Prison Brief, 2021
[22] As noted above it is beyond my scope to examine the instrumental value of moral responsibility/desert

preservation of the concept of moral responsibility after desert's denial, disproving desert would clearly necessitate a significant revision of our practice and beliefs.

MORAL RESPONSIBILITY SKEPTICISM

Many arguments against moral responsibility work by attempting to establish that we lack free will, widely considered a necessary property for moral responsibility. The two most notable exceptions are arguments that 'moral luck' precludes moral responsibility[23] and those that argue humans must be *causa sui* in order to be morally responsible.[24] My defence of moral responsibility skepticism will focus on Galen Strawson's framing of the latter *causa sui* approach.[25] By establishing 'being a *causa sui'* as a necessary condition for moral responsibility it is easier to see how free will, as a separate necessary property, might emerge unscathed from the denial of moral responsibility. This chapter serves as a blueprint for a more rigorous defence of the 'basic argument'.

STRAWSON'S BASIC ARGUMENT

These are the core premises of Strawson's argument, adapted for clarity from Strawson:[26]

1. You do what you do because of the way you are, specifically your mental constitution (MC).
2. To be morally responsible for an act at time T_i you must be responsible for having MC_i at that time.
3. To be morally responsible for being MC_i, you must have intentionally brought it about at some prior time (T_{i+1}). This is another act you may or may not be morally responsible for.
4. Moral responsibility requires completion of the infinite regress formed by 2 and 3.
5. Humans are finite, non-*causa-sui* beings.
6. From 4 and 5, you can never be morally responsible for what you do.

We make the choices we do through some combination of our beliefs, desires, preferences, instincts, rational deliberations etc. Whatever the particular account of action, it seems *prima facie* natural that to be responsible for the action we must be responsible for the aspect of ourselves that produced it. For example, if somebody were hypnotised so that they would feel an overwhelming desire for violence when they saw the colour blue, this would affect whether we ascribe moral responsibility for subsequent actions they take (even though they genuinely desire to do them). In mundane contexts we tend to assume that people *are* morally responsible, for both their actions and their character. Strawson tries to show that this is an error. Our actions are due to characteristics that were ultimately down to some combination of our environment and our genetics, neither of which are up to us. According to the 'control principle', 'A person's praiseworthiness and blameworthiness is restricted to what is within her control'.[27] This

[23] Nagel, 2012; Williams, 1981; Levy, 2011

[24] Nietzsche, 1886/1966; de Spinoza, 1677/2021

[25] The Impossibility of Moral Responsibility, 1994

[26] 1994, pp. 13-14

[27] Hartman, 2018, p. 170

captures why the hypnotised person isn't morally responsible according to folk intuitions and nobody is under Strawson's analysis. We can change ourselves, but only based on already existing preferences/characteristics and more external influences. If the reason somebody acted poorly was that they were selfish or weak willed or didn't care about others, then according to Strawson they have to be responsible for being that way to be responsible for the action. Nobody can be morally responsible because our MC at any point is the product of previous action/inaction that itself occurred due to a previous MC, and so on until we reach our universal starting point as a non-responsible baby.

One classic criticism of the basic argument, from Fischer, claims that it is overly demanding. According to Fischer, the agent is expected to have 'total control' over all factors that contribute to bringing about the action.[28] This clearly *would* be too demanding, requiring that the agent be responsible for the fact they had lungs and that the Sun gives off enough heat for life. As Michael Istvan points out, however, Strawson demands far less. The basic argument claims that the agent is not ultimately responsible for *any* of the factors that contributed to their action. So even on an account where moral responsibility only required responsibility over some 'infinitesimal bit of [an action's] sufficient cause', the basic argument would succeed.[29]

In 'Constitutional Responsibility' I show that to be responsible for even an 'infinitesimal' bit of our MC requires a defence of 'transcendental responsibility', i.e. an account of how we can be responsible for an action when we are not morally responsible for *any* part of our MC that brings it about.

Response Type I – Constitutional Responsibility

One way of establishing constitutional responsibility without transcendence is to deny premise 5, claiming that humans are either *causa sui* or non-finite beings. The former fails because the idea that something could cause its own existence is, as Nietzsche put it, 'the best self-contradiction that has been conceived so far, it is a sort of … perversion of logic'.[30] To cause something requires existence, so something would have to exist before its existence. Similarly, even if somebody wanted to defend the claim that humans aren't finite 'forever avoiding checkmate is not the same as winning'.[31] For many people infinite conceptions of human life are absurd, so I only briefly touch on them, but for a more detailed discussion of why even granting these extreme properties does not help see Istvan pp. 410-415.

A popular group of approaches, which Susan Wolf calls 'deep-self' views,[32] also fail to establish constitutional responsibility. The key idea is to distinguish our 'true' or 'deep' self in order to separate actions genuinely endorsed by us from those, for example, driven purely by desires we don't identify with. This is clearly an important aspect of our free agency. As Wolf points out 'It explains why kleptomaniacs, victims of brainwashing, and people acting under posthypnotic suggestion may not be responsible for their actions'.[33] Their actions are driven by compulsions and so, like somebody's actions when held at gunpoint, only 'theirs' in a minimal sense of the term. The issue is that, whilst it reveals a lack of freedom in these types of people,

[28] 2006, p. 116
[29] Istvan Jr., 2011, p. 403
[30] 1886/1966, §21
[31] Istvan Jr., 2011, p. 411
[32] Wolf, 1988; Frankfurt, 1971; Watson, Free Agency, 1975
[33] Wolf, 1988, p. 50

it fails to establish moral responsibility for the rest of us. It only provides answers, as Waller puts it, for the 'internal' debate of moral responsibility.[34] That is, assuming that humans can be held morally responsible, who in particular should or shouldn't be. The vital external question is not satisfactorily addressed. The fact that your deeper self affirms an action only grounds moral responsibility if you are responsible for that deeper self. The 'hypnotism' case can simply be replaced with Waller's 'evil potion' case, where somebody is turned vicious to the deepest level of their self.[35] Clearly it is still relevant to moral responsibility *how* people got the self they currently have.[36]

To really lock in the intuition behind the basic argument and see why at least some account of transcendence is required, it is helpful to imagine following a human from birth.[37] A baby is clearly not responsible for its MC as it had no control over the genetics and initial environment that shaped it. This is true from the shallowest to the deepest aspects of their self. Any of the earliest decisions a human makes will be based on the MC that they already possess and are not yet morally responsible for. For them to become responsible for their constitution at some point in their lifetime they will either have to perform an action that they are responsible for, despite not being responsible for their constitution at the time, or be responsible for the consequences of an action, despite not being responsible for the action. The latter directly contradicts what it means to be responsible for an action. The former is what I term 'transcendental responsibility'. Either way, it is a sort of moral responsibility equivalent to abiogenesis, which I dub 'aresgenesis'. Abiogenesis being the emergence of life from inanimate matter and aresgenesis the parallel transition from an *entirely non-responsible* self to an at-least-partially-responsible self.

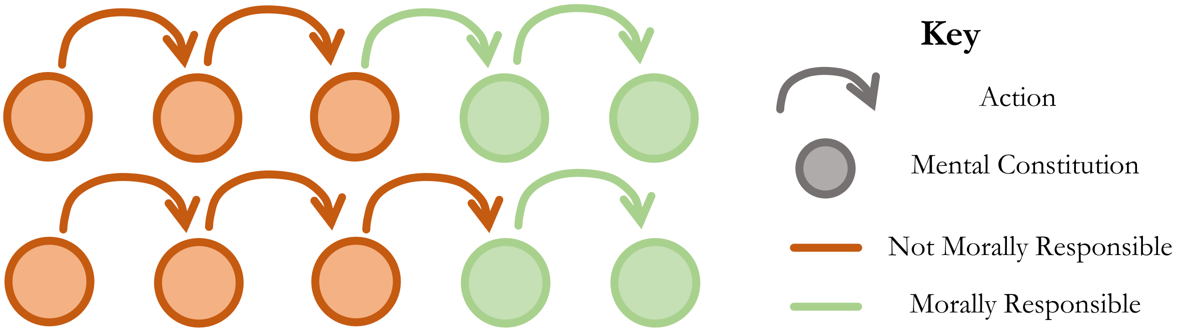


Figure 1 Aresgenesis

Response Type II – Transcendental Responsibility

To deny premise 2, accounts must establish why it is fair to hold somebody morally responsible for an action that emerges from a mental constitution that they aren't in the slightest responsible for. In this section I sketch how any such transcendental account can be refuted.

[34] 2011, p. 3
[35] 2011, p. 65
[36] This is not to claim that deep-self views don't address this, for example Dworkin requires 'procedural independence' and the ability to change first-order motivations for responsibility (1988, pp. 63,64), but appeal to the deeper self alone is insufficient. My arguments against transcendence should equally apply to deep-self views. For a more targeted rebuttal see (Waller, 2011, pp. 59-74)
[37] Strawson G. , 1994, p. 7

One compatibilist claim is that we can ‘shape’ ourselves through human capacities such as ‘reasons-responsiveness’, and it is this that allows us to transcend our constitution. Intuitively we do feel responsible for our actions and current character, without believing ourselves to be some strange ‘uncaused cause’. Most of us have made considered choices that helped shape us into the people we are today. This ability to self-shape, if not self-create, seems *prima facie* sufficient to make us morally responsible for our selves. With some minimum of reasons-responsiveness agents should have an awareness of right and wrong. Even though humans start with no constitutional responsibility, these capacities seem to allow somebody to be at least marginally responsible for their actions. Once even a little responsibility is granted, ‘the more of these actions that she performs for which she is truly morally responsible to some degree, the more she becomes truly morally responsible for who she is’.[38] This responsibility for the self accumulates till they can be considered fully responsible agents. This nicely fits with our intuitions. For example, the degree that we hold people responsible does seem to operate on a continuum from a baby to a young child to a teenager to a fully rational adult. Despite our initial impression, however, this does not justify moral responsibility.

Compatibilist ‘reasons-responsiveness’ accounts fail because it is unfair to ascribe even a tiny amount of responsibility when the person is not responsible for *any* aspect of their MC. Whilst it is not exhaustive, Sripada points to two structures of reasoning on which people could base these ‘self-changing’ and supposedly ‘minimally responsible’ actions.[39] On the first reasons are grounded in existing attitudes of the self. This is intuitive, as we do have subjective desires and beliefs that appear to generate the reasons for our actions. However, as Sripada points out, ‘The problem is that the kind of control it envisions is excessively self-ratifying and can’t be used to fundamentally change one’s existing self'.[40] During a child’s ‘aresgenesis action’, all of her priorities, motivations, and reasons are aspects of her MC, for which she is not yet responsible. To act differently she would need to already have different ideas and motivations, but as this is her aresgenesis action she is definitionally not responsible for any previous action that might have brought a different MC.

The second approach is ‘selection by objective reasons’. According to this, ‘regardless of what attitudes a person currently has, so long as she has a basic understanding of morals … she can nonetheless recognize the reasons ‘that there are’ to be different’.[41] This has the advantage that the same reasons are accessible to anybody with a minimum of rationality, giving us a sort of ‘moral equality of opportunity’. The problem is that under many popular accounts, attitude independent reasons don’t have a necessary link to motivation. So ‘a person will change her self based on recognition of objective reasons to change only if she antecedently cares about those kinds of reasons’.[42] Again, if the child does not act on the objective moral reasons because she does not care about moral reasons, that is not her fault. She is not yet responsible for the fact that she doesn’t care. Even if objective reasons *were* intrinsically bound with motivation it would presumably provide the same normative force to all people and so not relevantly explain differences in action between people.[43] Other, more

[38] Hartman, p. 175
[39] Sripada, 2017, p. 15
[40] Ibid., p. 15
[41] Ibid, pp. 15-16
[42] Ibid.
[43] Either people would have to be aware of different reasons, which would not be their fault, or the difference would be random or still explained by their MC

complex, accounts might be put forward, but the general problem can be seen.[44] Without at least some prior responsibility for their existing MC it appears unfair to hold somebody even minimally responsible for choices determined by cares and reasons they are not at all responsible for.

It is only by generalising and avoiding the particular agent and the process by which they make a choice that transcendent responsibility can seem possible for determinists. When we actually look at why they acted as they did, even if it reveals a weakness or moral flaw, it is clear they would have had to *be* different to act different. I think the compatibilist 'self-shaped' human only seems to work because we don't actually picture somebody not *at all* responsible for their MC at the start. We can imagine growth in moral responsibility from a tiny seed, but the change from not-at-all responsible to 0.000001% responsible for your MC is infinitely bigger than that from 0.000001% to 90% responsible.

Kane's event-causal libertarianism fails to avoid the above hurdle. According to Kane, moral responsibility can be rescued by the fact that the world does not actually appear determinate on current scientific models. Strawson points out, however, that it's unclear how Kane's random indeterminacy would add responsibility.[45] If I am not responsible for which options I am choosing between, flipping a coin to decide would not somehow make me responsible. Kane's event-causal libertarianism is just as susceptible as the compatibilist accounts above.[46]

The final accounts I will frame a Strawsonian attack against are agent-causal libertarian accounts. For many these views can be rejected out of hand for the unnaturalistic properties they pose. They claim, something like, that humans are in part constituted by an agent self that is not constrained by regular laws of causation. Instead, the agent is the immediate cause of their actions. Each human has 'a prerogative which some would attribute only to God: each of us, when we act, is a prime mover unmoved'.[47] I believe that humans do not have these properties, but Strawson's argument can potentially be shown to succeed even when they are granted. Should this fail I will simply fall back among the ranks of compatibilists and hard incompatibilists that deny we have these powers.

Most of the libertarian accounts having something like the person's physical MC, subjective or objective reasons, and an agent-self that chooses between them. 'For Pereboom, the agent-self is what brings A about, and in fact does so by acting on the motive that is the strongest *according to its own estimation*'.[48] Coming to our 'aresgenesis action', the person is in no way responsible for their physical MC, so responsibility must come from the input of the agent-self alone. Istvan claims that 'since it acts to bring A about, it must have some MC based on which it acts'.[49] If it were characterless its causal input would either be totally random or impotent. At the very least, 'if in conceiving the self you detach it from all motives or tendencies, what you have is not a morally admirable or condemnable, not a morally characterisable self at all'.[50] We can allow that humans have the ability to do otherwise on the 'categorical analysis', according to which 'an agent S has the ability to choose or do otherwise

[44] For more detail see (Sripada, pp. 14-17)
[45] 1994, p. 18
[46] For more detailed arguments see (Waller, 2011, pp. 35-38; Istvan Jr., 2011, pp. 415-417)
[47] Chisholm, 1964, p. 12
[48] Istvan Jr., p. 408
[49] Ibid., p. 411
[50] Hobart, 1934, p. 5

than ϕ at time **t** if and only if it was possible, holding fixed everything up to **t**, that S choose or do otherwise than ϕ at **t**'.[51] Even if they can choose otherwise, the way they choose is always going to be grounded in who they are. That is why we consider it *their* action, one based on which we can ascribe moral judgements in the first place. In effect I agree with Hobart's analysis of the libertarian that 'he is simply setting up one character within another'.[52] The agent-self is subject to the same analysis of reasons from Sripada and still falls to the basic argument. For a more detailed response to Mele, Pereboom, and Clarke see Istvan Jr., pp. 406-415.

Though obviously not conclusive, this section demonstrated how Strawson's basic argument can be brought to bear against standard defences of moral responsibility. Any account of moral responsibility will need some account of moral 'aresgenesis'. Most accounts were argued to fail at this. The next section argues that, despite not being morally responsible, we should consider ourselves to have free will.

Free Will Preservationism

> 'To know all is to forgive all.' True or false, that claim is not frightening. Some may regard it as naive, and others will view it as implausible or even foolish. But few will find it frightening. 'No one has free will.' That is a more chilling thought altogether.[53]

Free will is near-universally accepted to be closely related to moral responsibility. For many this is a relationship of necessity, for others it is one of sufficiency. For some, free will is in fact *defined* as the agential control required for moral responsibility.[54] Arguing for free will's existence alongside a denial of moral responsibility is therefore deeply radical. This alone is not enough to rule the position out, however. As Waller points out, changes in human understanding of themselves have long been a valuable project "from Copernicus to Darwin to the present, and those changes do not mean that Copernicus was no longer talking about the Earth, that Darwin had abandoned discussion of the human species" (2015, p. 189). Further, our denial of moral responsibility relies on establishing a second necessary property for moral responsibility; being a *causa* sui. This rules out the possibility of free will being *sufficient* for moral responsibility but is still compatible with a kind of free will that is *necessary* for moral responsibility.

This section argues that we should consider ourselves to have free will even after accepting the arguments against moral responsibility outlined above.

Preservationism, Eliminativism, & Reference I

In 'Free Will and Error', Shaun Nichols argues that when a concept is shown to involve an error it remains *prima facie* open whether we should respond as an 'eliminativist' or a 'preservationist'. In the case of free will, 'the eliminativist … maintains that there is *no such thing as* free will and that everyone is under the illusion that there is free will. The preservationist … says that there *is* free will and that everyone has merely been under some

[51] O'Conner, 2021

[52] Hobart, 1934, p. 5

[53] Waller, 2015, p. 199

[54] O'Conner, 2021

misapprehensions about its nature'.[55] To demonstrate the viability of both responses, Nichols compares two historical instances of conceptual error. Upon discovering that 'Whales' were not in fact fish - a previously assumed property - the concept was retained and revised, considered to successfully refer to a real, sea-dwelling creature despite the error.[56] By contrast the concept of 'phlogiston', a theorised flame-like element released by combustion, was eliminated. The degree of its error was such that it was considered to fail to refer altogether.[57] This section takes Nichols's work, and Caruso's response in 'Free will eliminativism: reference, error, and phenomenology', as its starting point. This amounts to accepting Caruso's claim that 'all such debates boil down to whether or not the erroneous folk term in question successfully refers or not'[58] and investigating whether free will without moral responsibility and the categorical ability to do otherwise (CADO)[59] should be considered to refer.

According to the causal-historical account of reference, 'words refer in virtue of being associated with chains of use leading back to an initiating use or 'baptism' of the referent'.[60] In this way Nichols suggests preservationists might claim that free will continues to refer despite significant error, in our case attribution of moral responsibility and the CADO. However, Caruso argues that even on a causal-historical account reference would fail as 'it is *prima facie* plausible to think that the concept of 'free will' was originally baptized in a causal-historical story that appealed, at least in some significant way, to our first-person experience of free agency'.[61] The link to first-person experiences of free agency seems right, but Caruso's inference that it therefore fails to refer is not so clear. To show that 'free will' still refers on the casual-historical account will require demonstrating that this 'first-person experience of free agency' 'baptised' something that survives the denial of moral responsibility and the CADO.

In descriptive accounts of reference a word refers in virtue of its association with descriptive content which 'picks something out' as the referent.[62] According to Caruso, 'On a descriptive account of reference, eliminativism follows since, as Nichols himself argues, the folk concept of free will contains significant error, hence nothing satisfies the description'.[63] One way to argue for this would be to claim that free will's descriptive content includes the idea that it is sufficient for moral responsibility. If, as I have argued, it is an error to claim that humans can be morally responsible, free will would fail to refer to anything at all. However, that we have moral responsibility seems more like an obvious *consequence* of having free will than a *component* of the concept itself, which *prima facie* is simply about how free our will is. As the only widely acknowledged necessary condition for moral responsibility, free will naturally *appears* sufficient for moral responsibility, but the whole point of the previous chapter was to reveal a second necessary condition; that of being *causa sui*. As an analogy, picture a teacher who assumes that a genius student will pass an exam but is wrong because there is a physical fitness component they were unaware of. The fact the student doesn't pass the test doesn't then mean the teacher's belief in their intelligence was incorrect, the teacher's misunderstanding of the test to be met is to blame. Equivalently, humans failing to meet the

[55] 2013, p. 209
[56] Ibid., p. 208
[57] Ibid., p. 204
[58] 2015, p. 2824
[59] See pg. 11
[60] Michaelson, 2021
[61] 2015, p. 2828
[62] Michaelson, 2021
[63] 2015, p. 2824

criteria for moral responsibility indicates a misunderstanding of its requirements, not that we must lack free will.

To demonstrate that 'free will' continues to refer on the descriptive account I will therefore argue that its descriptive content picks out agential capacities that we still possess and none that we lack. This allows me to stay more firmly rooted in the literature, where most of the discussion is about whether we can have free will without the CADO (which we also lack under my account).

The subsequent sections discuss the agential capacities we still possess on this account and the folk conception of 'free will', arguing that taken together they suggest that 'free will' continues to refer on both accounts of reference.

NATURALISTIC FREE WILL

In determining whether we have free will it is helpful to start with what we might pre-theoretically expect the term 'free will' to describe. To this end its hard to imagine something less controversial than Robert Hobart's claim that 'free will means freedom of persons in willing'.[64] In what ways do humans 'will' things and do they do so 'freely'? To address the former, we can turn to some of the compatibilist accounts argued against above.

Analyses of deep-self views, such as Frankfurt's and Wolf's, appear to pick out important properties of our will. There appear to be different 'levels' of our selves, willing independently on different bases and, potentially, contradicting each other. For example, we might be tired and so at a 'lower level' desire to just lie on the sofa, whilst on a 'higher' one we wish to 'make something out of every day' and 'get out into nature more'. It also seems right that we identify more with certain levels than others. This is made clear by the fact that when the lower desire wins out we often feel that we 'succumbed' to what we didn't 'really' want. By contrast when the higher desire wins out it feels more like a successful feat of our will, one of self-mastery. Humans are also 'capable of wanting to be different, in their preferences and purposes, from what they are'.[65] This links nicely to our capacity for reasons-responsiveness and self-shaping. The will can impose itself on the external world, for example a thirsty person has the ability to move their body to drink and quench that thirst. The hierarchy of a human self also allows them to impose their will on their own self. Upon noticing a tendency for tired laziness, and feeling there is reason to change this, the higher self might decide to establish a sleeping pattern in order to reshape these 'lower' desires. What we 'will' is evidently related in some way to what we desire, perhaps specifically at the level of our selves we identify with. Further, the process of 'willing' something to happen it appears to combine this preference with an active attempt to bring that end about.

In deciding to what degree the will is free, we can again turn to Hobart, 'The freedom of anyone surely always implies his possession of a power, and means the absence of any interference (whether taking the form of restraint or constraint) with his exercise of that power'.[66] This fits with Waller in *Restorative Free Will* where he argues that 'open alternatives' and 'control' are key for free will.[67] A person's will would then be free in as much as they can decide their own actions between genuine alternatives. This would certainly align

[64] Hobart, p. 8

[65] Frankfurt, 1971, p. 7

[66] Hobart, 1934, p. 8

[67] 2015, pp. 122, 132

with when we might feel we *lack* freedom in willing, for example if we were to will one thing but do another (such as involuntary movements from Tourette's Syndrome). Similarly, a lack of open alternatives can 'constrain' our will as easily as taking away our control. It also aligns with when we *feel* free, when multiple options lie open before us and we make the choice between them – based on our preferences or some external reason. In the chapters "Psychological Free Will" and "Restorative Free Will", Waller surveys an array of psychological research in which the importance of control and open alternatives for psychological wellbeing is apparent.[68] Waller uses similar studies to pull out internal features that can help or hinder the freedom of our will. For example, the strength of our 'sense of self-efficacy' and 'internal locus of control'.[69] These influence the range of options that appear genuinely open to us, as well as facilitate our ability to act as we will against internal or external resistance. Under this view, freedom of the will is a complex but natural amalgamation of psychological and situational features pertaining to how our will can freely roam over open alternatives and exercise control over its future.

The above picture of naturalistic free will lacks a historically/philosophically significant feature: the categorical ability to do otherwise (CADO). I have been claiming that humans' naturalistic free will is grounded in 'open alternatives' and 'control', but these are precisely what incompatibilists want to put pressure on. If it is built into the laws of the universe that I will make a certain choice, then in what way are the alternatives open to me? In what way am I in control? The naturalist would respond that we are made up of our psychology, our MC, desires, thought processes etc. We don't choose this self in a way required for ultimate responsibility, but that does not mean that our self is not us. When we act predictably off our thoughts and desires, they are *our* thoughts and desires. Sometimes desires can seem to overrule what we think of as the 'real' us, for example cravings from addiction, but deep-self frameworks reveal how to separate these from cases where we identify with our choices and reflectively endorse them. 'It is no proof that I cannot do something to point out that I shall not do it if I do not prefer'.[70] We have evolved to explore and select between open alternatives and this has great psychological value for us, but in order to justify moral responsibility 'at some point the genuine need for open alternatives became transmogrified (or apotheosized) into a desire for absolute first cause miraculous choices'.[71] I make a fuller argument that the CADO is not in fact part of our folk conception of free will in the next section, but for present purposes the above will suffice.

According to Caruso, for 'free will' to refer on the causal-historical account we are looking for what was 'baptised' by our first-person phenomenology of free-agency. The answer is the capacities discussed above. We make choices, function with reasons-responsiveness, can make evaluative judgements along many different axes, and override lower-level desires based on higher-level preferences that we more deeply identify with. That we lack the CADO appears to better fit the preservationist claim that 'everyone has merely been under some misapprehensions about its nature'[72] than that the concept doesn't pick out anything at all. Much like the 'whale' example, the freedom of our will largely fits our conception beyond one capacity that, I will argue in the following subchapter, is metaphysically odd, not particularly

[68] Ibid., pp. 115-168
[69] Ibid., p. 134
[70] Hobart, 1934, p. 11
[71] Waller, 2015, p. 141
[72] Nichols, 2013, p. 209

desirable, and not actually part of our folk conception anyway. The sentiment of our freedom of choice is grounded in desires, preferences, and rational considerations, all of which I argue survive. So 'free will' continues to refer on the causal historical account, just as Nichols concludes.

The next section argues that the CADO is not in fact part of our folk conception of free will, meaning that free will continues to refer on the descriptivist view.

Sinister Laplacean Demon or Loving Molinist God

Laplace's 'demon' is one of the earliest and most compelling framings of the problem generated for free will by determinism. Laplace describes 'an intelligence which could comprehend all the forces by which nature is animated and the respective positions of the beings which compose it … for it, nothing would be uncertain and the future, as the past, would be present to its eyes'.[73] The problem for free will can be laid out like so:

1. A being (X) with free will can choose freely between two actions A and B
2. Laplace's demon knows the positions, energy states etc. of X's atoms before their 'choice'
3. From 2, and its complete knowledge of the deterministic laws of physics[74], the demon knows the positions of X's atoms after they make their 'choice'
4. From 3 the demon knows X will choose A
5. The demon cannot be wrong as humans lack the ability to break the laws of physics and causation
6. From 4 and 5, X will choose A and *cannot* choose B

C. From 1 and 6, determinism and free will cannot co-exist

Under the eyes of the demon, actions that had appeared freely chosen reveal themselves to be the necessary unfolding of cause and effect under the laws of physics. Framed in this way it is hard to see how we could maintain a sense of freedom because they lack what is recognisably the CADO. The human appears to lack control over their choices and there seemingly aren't any open alternatives.

To ground the compatibilist intuition of freedom in a determined world we can look to Molinism, put forward by 16th century Jesuit theologian Luis de Molina. Its purpose is to resolve the apparent conflict between God's omniscience and humanity's free will. In his recent work on Molinism, Ken Perszyk notes that 'The Molinist account of God's providence … seems functionally equivalent to Laplace's 'intelligence''.[75] The Molinist's problem, and its similarity to the determinist's, is revealed by the argument below:

i. A being (X) with free will can choose freely between two actions A and B
ii. God is infallibly omniscient (all-knowing)
iii. From 2, God knows X will choose A

[73] Laplace, 1902/1814, p. 4
[74] For our purposes we can ignore indeterministic features given, as noted against Kane, they do not appear to aid freedom and/or control
[75] 2000, p. 16

iv. God cannot be wrong since an infallible, omniscient being cannot have false knowledge

v. From 3 and 4, X will choose A and *cannot* choose B

C. From 1 and 5, omniscience and free will cannot co-exist

The Molinist response revolves around specifying the type of knowledge that God has about human actions. They claim that God has ‘middle knowledge’, which is knowledge of ‘counterfactuals of freedom whose truth-values are contingent but *independent* of God’s will’.[76] That is to say that God’s middle knowledge reveals what we would *freely choose* to do in any given circumstance. God’s middle knowledge that X will choose A is held to be analogous to somebody who knows their friend so well that they know, for example, what drink they will freely choose. My knowing somebody’s character, and so how they are likely to act, doesn’t limit their freedom. The key difference is that God’s omniscience means they know X maximally well. They not only know X’s general preferences but all of their recent thoughts, what drinks they’ve been meaning to try, what sorts of new drinks would attract them, the exact mood they’re currently in, how spontaneous they are… and so on. The Molinist allows that there is a matter of fact how we *will* act in a certain situation, whilst denying that this should prevent us from claiming that we can choose as we wish. So they deny premise 5, claiming that ‘X will choose A’ simply means that ‘X *will not* choose B’ rather than ‘*cannot’*. Both options were equally open to them but given their character and state of mind at the point of choosing there is a fact of what they would select. The fact that an all-knowing, all-loving God knows how they will act is held up as a testament to sheer scale of God’s knowledge and love, rather than an indication of the limitations of their freedom.

Under this framework, the CADO is the ability to ‘act as we wouldn’t’ in a given situation. Described in this way it appears an absurd capacity to desire. It is the ability to act so randomly and unpredictably that no matter how well somebody knew us, including a God, it couldn’t be known beforehand. To what degree would this kind of action stem from something we would recognise as ourselves? Our concept of self is inextricably linked to our personalities, temperaments, habits, preferences etc. Our ‘unknowable’ free action would feel less like a radically free moment where we subverted God’s plan and more like our body had been possessed and moved against our preferences, one of the least free feeling things conceivable.

We might now understand how a Molinist could feel free with their all-knowing and all-loving God, and with actions that were ‘inevitably’ going to turn out the way they did, but it remains unclear how this would help the determinist. The determinist is stuck with a metaphorical demon which doesn’t have to love or even ‘know’ them at all. As highlighted in the consequence argument, my actions are wholly decided upon and predicted through physics at the level of my atoms.

According to liberal naturalistic views of reality, there are distinct descriptive levels of reality. Reality at a ‘higher level’ of description can have properties that do not operate at lower levels, for example the laws of thermodynamics or human emotion. According to Sean Carroll, the mistake made when it comes to determinism and free will is the linguistic movement between these distinct levels of description.[77] To start off with a description of atomic position and end with a conclusion about human choice is deeply misleading. For example, it would not seem legitimate to answer the question “why was that joke funny?” with a series of equations

[76] Ibid., p. 14

[77] Carroll, Poetic Naturalism Lecture, discussion from 35:30

and an excel spreadsheet full of atomic coordinates. 'Choice', 'decisions' and 'actions' are properties at the level of human agency, not at the level of atoms. On this view *explaining* a human's actions through physical laws is an illegitmate move. This move therefore effectively rejects premise 6. The demon knowing what X will do through its atomic knowledge does not mean that this information is *why* they do what they do. Nor does it mean they *cannot* do otherwise, simply that they *will* not.

In response we might query the legitimacy of this restriction. It does seem that if an agents choice was between moving in two different directions the atomic information would easily explain this. Further, in Carroll's account he does allow for one 'level of description' to be 'mapped' onto another. The incompatibilist might therefore claim that this is exactly what happens in the case of the demon. Given the position of X's future atoms it could be quite easy to translate to a result with respect to their choice. It is important here to distinguish prediction from explanation. From the aforementioned atomic information the demon *might* be able to show exactly how the auditory vibrations of the joke lead to the neurochemical changes that code for funniness, but this would still not be valid as an explanation of *why* it is funny at the agential level. Its just a demonstration *that it is* so. Similarly, all that determinism shows us is what an agent *will* do, which Molinist considerations allow us to see is unthreatening to our freedom. The incompatibilist is 'merely making a slip in the use of the word 'could''.[78] This can be further reinforced, as this map from atomic information to agential information would apply to all the demon's knowledge. That means that premise 2, the demon's complete knowledge of X's atomic makeup, maps onto the demon's complete knowledge of X's characteristics at the agential level. That is their memories, thoughts, desires, mood, preferences etc. The demon could be therefore considered to collapse into the Molinist God entirely.

The above analysis aligns with and resolves seemingly conflicting experimental data on folk intuitions surrounding free will. In his paper "Folk Intuitions on Free Will", Shaun Nichols examines whether people's intuitions are best characterised as deterministic or indeterministic. He summarizes his findings, 'When engaged in the practical process of predicting and explaining behavior, people treat choice as deterministic, but in other contexts, people seem to regard choice as indeterminist'.[79] This goes against common assumptions held in the literature that the folk conception is solely indeterministic.[80] According to Nichols's research, seemingly contradictory intuitions can be pulled out depending on how the situation is framed. All of the scenarios that delivered incompatibilist intuitions hinged on the question of whether, given fixed physically defined states of affairs, certain human decisions "*had to happen*".[81] By contrast, deterministic perspectives were produced when people were asked to predict the behaviour of individuals based on the actions of a constitutionally and situationally identical individual. In these contexts, the majority responded as determinists.

Under Nichols's analysis, these two responses produce a tension which he suggests is potentially what generates the free will problem in the first place. 'On the one hand, it seems plausible to us that our decisions flow directly from our psychology. On the other hand, it seems plausible to us that our decisions are not simply a consequence of past forces'.[82] It might

[78] Hobart, 1934, p. 9
[79] Nichols, 2006, p. 65
[80] Ibid., p. 59
[81] Ibid., pp. 66-67
[82] Ibid., p. 73

be considered a necessary criterion for a successful account of free will that it resolves this tension, for example by providing convincing reasons for denying one of the conflicting intuitions. Our insights from Molinism and liberal naturalism, by contrast, allow us to understand *both* intuitions and resolve the apparent conflict between them. Our sentiment of freedom is grounded in our making decisions based on our character, beliefs, and desires. When descriptions remain grounded in this agential level, we can accept that there is a matter of fact of how we *will* act given the totality of our state of mind. Our seemingly-indeterminist intuitions are only brought about when this fact is framed as external laws *forcing* separate agents to act that way. When we properly understand the distinct physical and agential levels of description, however, we should know that there are no agents to be forced to do anything at the physics level of description, just more atoms and energy states. The mistake is in blending these two levels of description and so getting something as unintuitive as somebody's desires shaping the laws of physics. Unless we want to deny the legitimacy of absolutely everything at any level of description other than that outlined in physics – a pretty extreme and unattractive position – we should recognise the laws of physics as *constraints* on higher levels of descriptions but not *explanatory* of them. Saying we lack free will because there is no wriggle-room for freedom in the laws of physics is like arguing a joke isn't funny because the laws of physics aren't humorous. The seeming conflict is produced by an error, but it is an error on the part of the philosophers who established the cases.

Based on the arguments above, I believe that 'free will' still refers on the descriptive account, even with the denial of moral responsibility and the categorical ability to do otherwise. When appropriately described our folk conception of free will is compatibilist. It is the linguistic movement between atomic and agential levels of description that leads to seemingly incompatibilist responses.

PRESERVATIONISM, ELIMINATIVISM, & REFERENCE II

Even if above argument fails, so one account of reference justifies preservationism and the other eliminativism, the question would remain which view of reference is apt. Nichols adopts a sort of pluralism where, depending on the particular case, the account of reference which makes sense to apply changes. For example, if we found out that whales were actually Martian spy robots it might no longer be enough that the term points to the same sea-dwelling things to think they're still 'whales'. Similarly, Nichols highlights historical cases where discussions of phlogiston would interchangeably operate casual-historically *and* descriptively.[83]

According to Nichols, 'we might appeal to practical interests in deciding which convention to adopt and impose'.[84] He lists pros of eliminativism and preservationism. For preservation Nichols points to a growing body of experimental work that suggests that telling people they don't have free will has negative social consequences, for example a marked decrease in pro-social behaviour.[85] For elimination Nichols looks at the negative consequences of free will belief, such as retributive attitudes, psychologically crippling guilt, and lack of compassion for people with 'bad' qualities.[86] However, the specific 'error' that Nichols is considering is the apparent intuition of indeterminism, whereas we are looking at RSC, so our

[83] 2013, p. 210
[84] Ibid., p. 215
[85] Ibid.
[86] Ibid., pp. 215-216

conceptions of what would be retained with ‘free will’ is different. As Waller points out, ‘Once we separate free will from moral responsibility, Nichols’ scorecard for the practical advantages of keeping vs. eliminating free will turns out to be: Advantages for keeping free will, many; advantages for eliminating free will, zero’.[87] The benefits come from a sense of forward-looking agency and control, an ability to shape our lives based on our preferences and make decisions that we reflectively endorse and live up to our standards for ourselves. Simultaneously we remove backwards looking attitudes that aim to respond to wrongdoing with ‘righteous suffering’. If Nichols is right that the matter should be settled on practical grounds, then the RSC position appears to be the best of both worlds.

Caruso denies the practical benefits of this position, claiming that ‘By liberating free will from moral responsibility, Waller has seemingly liberated it from all of its philosophical and practical importance’.[88] This misses the immense importance of free will for our sense of self and wellbeing. For exposition on a range of psychological research demonstrating the extent of this, see Waller’s 2015, pp. 115-168.

So either through our Molinist framework, or on pragmatic considerations, we *should* liberate free will from moral responsibility. The agential capacities we possess, to varying degrees, are severely misrepresented by the claim that humans lack free will. To take a perfect summary passage from Waller’s *Restorative Free Will:*

> “You have genuine alternatives from which to choose, you can make your choices in accordance with your own values and preferences and understanding, and you can exercise genuine control over your own choices. You can’t perform miracles; but you never really believed that you could. You can’t make choices that are totally independent of who you are, and your character and values and history and circumstances; but if you could, it is hard to imagine that you would actually be making the choices, and difficult to see how ‘your choices’ would be of any benefit to you.” What else would you want in the way of free will?[89]

Conclusion

This dissertation has attempted to ‘unshackle’ free will from moral responsibility. I first defended the claim that the denial of desert-based moral responsibility constitutes the denial of everything worthy of the label ‘moral responsibility’. Alternatives miss something core to our intuitive conception of moral responsibility, often picking out people as morally responsible based on unacceptable features. I then outlined a defence of Strawson’s ‘Basic Argument’. As human’s are born with a mental constitution that is generated by genetic and environmental factors, and so is outside of their control, they begin lacking moral responsibility for their constitution. This means a moment of moral ‘aresgenesis’ is required. A viable defence of moral responsibility therefore needs to justify blaming somebody to some degree for a choice when they are not responsible for *any* aspect of the self that makes that choice. Whether their reasons are objective or subjective, Sripada’s argument shows that no determinist account could succeed. Libertarians were also argued to fail. Having shown how moral responsibility can be excised without appeal to free will, I finally argued that what remains of our human capacities includes what should be deemed ‘free will’. Our agential capacities were argued to

[87] 2015, p. 198
[88] Caruso, 2016
[89] 2015, p. 200

be sufficient to ground the baptism of ‘free will’ on a causal-historical account of reference. Further, intuitions for the libertarian ‘categorical ability to do otherwise’ were argued to rely on a misleading substitution of ‘can’ for ‘will’ and inappropriate blending of physics and human ‘levels of description’ leading to a false intuition pump. Ultimately free will was claimed to still refer, either on both accounts of reference or at least on pragmatic grounds.

Reverse Semi-Compatibilism is clearly a viable and attractive position within the free will debate. As Waller says, ‘there is really nothing new in this position; to the contrary, it simply combines elements of two positions that are widely held among philosophers’.[90] If the seeming ambivalence between incompatibilism and compatibilism can be resolved within one position, we might have an excellent case of Wittgensteinian ‘therapeutic philosophy’. Pessimists such as van Inwagen worry that without the illusion of moral responsibility we would be in dire straits. Empirical evidence from less punitive nations, such as Norway, imply otherwise. Rehabilitation and addressing social causes appear far more effective.[91] Further, Buddhist practice demonstrate the ability to incorporate these intuitions even into our immediate reactive emotions.[92]

The sooner that we are rid of the shackles of moral responsibility, the better, and the freer we can become.

References

Carroll, S. (2021, April 21). *Poetic Naturalism Lecture.* Retrieved from Preposterous Universe: https://www.preposterousuniverse.com/videos

Caruso, G. D. (2015). Free Will Eliminativism: Reference, Error, and Phenomenology. *Philosophical Studies*, 2823-2833.

Caruso, G. D. (2016). Free Will Skepticism and Criminal Behavior: A Public Health-Quarantine Model. *Southwest Philosophy Review*, 25-48.

Caruso, G. D. (2016). Restorative free will: back to the biological base. *Notre Dame Philosophical Reviews*. Retrieved from https://ndpr.nd.edu/reviews/restorative-free-will-back-to-the-biological-base/

Caruso, G. D. (2017). Public Health and Safety: The Social Determinants of Health and Criminal Behavior. *ResearchLinks Books*. Retrieved from https://ssrn.com/abstract=3054747

Chisholm, R. M. (1964). Human Freedom and the Self. *The Lindley Lecture*. University of Kansas. Retrieved from https://kuscholarworks.ku.edu/bitstream/handle/1808/12380/Human%20Freedom%20and%20the%20Self-1964.pdf?sequence=1

[90] 2011, p. 44

[91] Martinez, 2017, pp. 30-59

[92] Goodman, 2002

Clarke, R. M. (1997). On the Possibility of Rational Free Action. *Philosophical Studies*, 37-57.

Clarke, R. M. (2005). On an Argument for the Impossibility of Moral Responsibility. *Midwest Studies in Philosophy, 29*, 13-24.

de Spinoza, B. (1677/2021). *The Ethics.* Strelbytskyy Multimedia Publishing.

Dennett, D. (2021, April 19). *Exchange on Waller's "Against Moral Responsibility"*. Retrieved from Naturalism: https://www.naturalism.org/resources/book-reviews/dennett-review-of-against-moral-responsibility

Dworkin, G. (1988). *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

Fischer, J. M. (1998). *Responsibility and Control.* Cambridge University Press.

Fischer, J. M. (2006). The Cards That Are Dealt You. *The Journal of Ethics*, 107-129.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, 5-20.

Goodman, C. (2002). Resentment and Reality: Buddhism on Moral Responsibility. *American Philosophical Quarterly, 39*(4), 359-372.

Hartman, R. (2018). Constitutive Moral Luck and Strawson's Argument for the Impossibility of Moral Responsibility. *Journal of the American Philosophical Association*, 165-183.

Hobart, R. E. (1934). Free Will as Involving Determination and Inconceivable Without It. *Mind, 43*(169), 1-27.

Hume, D. (1748/2000). *An Enquiry Concerning Human Understanding.* Oxford: Clarendon Press.

Istvan Jr., M. A. (2011). Concerning the resilience of Galen Strawson's Basic Argument. *Philosophical Studies*, 399-420.

Kane, R. (2000). Responses to Bernard Berofsky, John Martin Fischer and Galen Strawson. *Philosophy and Phenomenological Research, 60*(1), 157-167.

Kant, I. (1790/1996). *The Critique of Judgment.* (W. S. Pluhar, Trans.) Indianapolis: Hackett.

Laplace, P. S. (1902/1814). *A Philosophical Essay on Probabilities.* London: John Willey & Sons.

Levy, N. (2011). *Hard Luck.* Oxford University Press.

Martinez, R. (2017). *Creating Freedom: The Lottery of Birth, the Illusion of Consent, and the Fight for Our Future.* Penguin.

Mele, A. R. (1995). *Autonomous Agents.* New York: Oxford University Press.

Michaelson, E. (2021, April 11). *Reference*. Retrieved from Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/reference/#FourMod

Nagel, T. (2012). Moral Luck. In T. Nagel, *Mortal Questions* (pp. 24-38). Cambridge University Press.

Nahmias, E. (2014). Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Volume 4: Free Will and Moral Responsibility* (pp. 1-57). London: The MIT Press.

Nichols, S. (2006). Folk Intuitions on Free Will. *Journal of Cognition and Culture, 6*(1-2), 57-86.

Nichols, S. (2007). After Incompatibilism: A Naturalistic Defense of the Reactive Attitudes. *Philosophical Perspectives, 21*, 405-428.

Nichols, S. (2013). Free Will and Error. In G. D. Caruso (Ed.), *Exploring the illusion of free will and moral responsibility.* Lanham: Lexington Books.

Nietzsche, F. (1886/1966). *Beyond Good and Evil.* (W. Kaufmann, Trans.) New York: Random House.

O'Conner, T. (2021, April 10). *Free Will*. Retrieved from Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/freewill/#FreeWillMoraResp

Pereboom, D. (2006). *Living Without Free Will.* Cambridge University Press.

Perszyk, K. (2000). Molinism and Compatibilism. *International Journal for Philosophy of Religion, 48*(1), 23.

Rosenberg, M. B., & Chopra, D. (1999/2015). *Nonviolent Communication: A Language of Life.* PuddleDancer Press.

Rottschaefer, W. A. (2014). Can We Responsibly Reject Moral Responsibility? *Behavior and Philosophy, 42*.

Scanlon, T. M. (1998). *What We Owe to Each Other.* Cambridge: Harvard University Press.

Schlick, M. (1930/1939). When is a Man Responsible? In *Problems of Ethics* (D. Rynin, Trans., pp. 141-158). New York: Prentice-Hall.

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics*, 602-632.

Smart, J. J. (1961). Free will, praise, and blame. *Mind*, 291-306.

Smilansky, S. (1994). The Ethical Advantages of Hard Determinism. *Philosophy and Phenomenological Research*, 355-363.

Smilansky, S. (2000). *Free Will and Illusion.* Oxford: Oxford University Press.

Smith, A. (2007). On being responsible and holding responsible. *Journal of Ethics*, 465-484.

Sripada, C. (2017). Frankfurt's Unwilling and Willing Addicts. *Mind*, 781-815.

Strawson, G. (1986). *Freedom and Belief.* New York: Oxford University Press.

Strawson, G. (1994). The Impossibility of Moral Responsibility. *Philosophical Studies,* 5-24.

Strawson, P. F. (1962/2003). Freedom and Resentment. In G. Watson (Ed.), *Free Will* (pp. 72-93). Oxford University Press.

Takakis, N., & Cohen, D. (Eds.). (2008). *Essays on Free Will and Moral Responsibility.* Cambridge Scholars Publishing.

van Inwagen, P. (1983). *An Essay on Free Will.* Oxford: Oxford University Press.

Waller, B. N. (2003). Denying Responsibility without Making Excuses. *American Philosophical Quarterly, 43*(1), 81-90.

Waller, B. N. (2007). Sincere Apology Without Moral Responsibility. *Social Theory and Practice, 33*(3), 441-465.

Waller, B. N. (2011). *Against Moral Responsibility.* MIT Press.

Waller, B. N. (2014). *The Stubborn System of Moral Responsibility.* The MIT Press.

Waller, B. N. (2015). *Restorative Free Will: Back to the Biological Base.* Lexington Books.

Waller, B. N. (2020). Beyond Moral Responsibility to a System that Works. *Neuroethics, 13*, 5-12.

Watson, G. (1975). Free Agency. *The Journal of Philosophy, 72*(8), 205-220.

Watson, G. (1996). Two Faces of Responsibility. *Philosophical Topics, 24*(2), 227-248.

Watson, G. (2018). Responsibility and the Limits of Evil: Variations on a Strawsonian Theme. In J. M. Fischer, & M. Ravizza (Eds.), *Perspectives on moral responsibility* (pp. 119-148). Ithaca: Cornell University Press.

Williams, B. (1981). Moral Luck. In B. Williams, *Moral Luck* (pp. 20-39). Cambridge University Press.

Wolf, S. (1988). Sanity and the Metaphysics of Responsibility. In F. Schoeman (Ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (pp. 46-62). Cambridge: Cambridge University Press.

World Prison Brief. (2021, April 20). *More than 10.35 million people are in prison around the world, new report shows*. Retrieved from World Prison Brief: https://www.prisonstudies.org/news/more-1035-million-people-are-prison-around-world-new-report-shows

Zimmerman, M. (1988). *An Essay on Moral Responsibility.* Totowa: Rowman & Littlefield.

Editors

External Reviewers