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/research/critique/

Address for enquiries and submissions:

durhamcritique@outlook.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

MMXXI **NEW SERIES** **Summer 2021**

CONTENTS

REVIEWS

The Death of Socrates, Jacques-Louis David, 1787

THE DEATH OF SOCRATES

In 399 BCE, Socrates, the founder proper of Western philosophy, was put on trial, charged with impiety (ἀσέβεια) and the corruption of the youth of Athens. Any Athenian citizen could bring forth this charge (ἀσεβείας γραφή) against any other, and it was mainly philosophers, and other intellectuals, that were charged with it by the Athenians, including, amongst others, Anaxagoras, Diagoras, Aristotle, Theophrastus and Protagoras.

The crime of impiety was ill-defined in Attic law, and therefore liable to become the pretext for persecution. Accusations were made for either breaches of the ceremonial law of public worship, or heresy, the former punitive of those who directly defiled the consecrated grounds and objects of the state religion, the latter intellectual in nature, often brought against those who directed not their swords but their words against the gods. Such charges are always more heavily grounded in personal sentiment, and the link between political alignments and the deployment of this broad charge is no coincidence.

The trial of Socrates was immortalized by Plato in the *Apology of Socrates* and by Xenophon in the *Apology of Socrates to the Jury*. However, it is not his trial but his death that is of special import to philosophers. Detailed in Plato's *Crito* and *Phaedo*, Socrates does not fear death, and does not therefore flee Athens, as many others charged with impiety did, and as his students implored him to, because he had faith in his philosophy of the immortality of the soul, and believed firmly in the democratic system of the Athenian state. The reason why philosophers have always held up Socrates as the paragon of philosophers is because, even in death, his actions met his convictions, so his words could not have been hollow. He drank the hemlock as though it were a draught of wine.

Editor's Introduction

THE POINT of an undergraduate journal of philosophy is to give students an opportunity to have published work that would not ordinarily be published in the competitive environment of academia. That is, undergraduate journals were designed specifically to afford papers a chance at publication that are not good enough to be worthy of publication. They separate out the upper echelons from the lower, precluding the former and inviting the latter to satisfy their own egotistical desires for recognition by winning a competition amongst the losers. Undergraduate journals are, generally speaking, anthologies of the best papers amongst the worst a field has to offer. This is why amateur means worse, and to call someone a professional is to assume a higher order of competence. There seems always to be a group of victors and a group of losers, and the latter group distinguishes itself into a new category, which it pretends is not lesser, so that they can succeed there where they could not succeed in the real event.

This is not the point of *Critique*. It is not commonplace for students to make original contributions to a field, whereas it is expected of professional scholars who make their living doing just that. Academic journals, however, are not simply for the purpose of publishing progressive works. Rather, they serve as antennae for a field too. We find many articles and journals focussing on matters of race and gender because they are the obsessions of the academics of our time, and we know that they are the obsessions of our time because we see the proliferation of articles and the creation of journals. The primary purpose of this journal, more so than the publication of exceptional undergraduate work that does indeed make some form of original contribution to the study of philosophy, is to serve as an antenna for undergraduate philosophic sentiment worldwide. If we should only ever read the work of previous generations, which is generally the matter of professional journals, we shall only know what old men think. To know what the young men are thinking – and we must not forget that the old must be at some point replaced by the young – we must read what the young are writing.

Unlike most journals, in which reviews and discussion pieces are peripheral or absent, this journal makes them the primary object of publication. We find that undergraduates are most suited to intelligent insights when they only have to make a few of them, and, when they have to write some several thousand words, not only does their capacity for intellectual excellence diminish, but their literary faculties suffer as a result of it too. Now it must be said that this is quite a generalization, and there are many exceptional students who write excellently and at great length, and many more professional philosophers which write drearily at any length, but it is quite fair to say that the most direct survey of any intellectual climate is through smaller expressions rather than the larger, else we should establish a publishing house instead, and solicit tomes from students rather than papers at all.

This issue of *Critique* is the first of its new series, which has been announced as a consequence of new editorship and a new philosophy of organization and intellectual survey, which has been synthesized with the founding philosophy of discourse. It is regrettable to admit that the journal's archive is in a state of disrepair, and a black mark on its history is the decline in its quality since the early years following its initial publication in June of 2009. The present Editor thought it wise to rebrand the journal with its rejuvenation, and, as a result, a new series was announced, a new website was established for the location of the publication of *Critique*, a new typesetting was introduced to the journal, submissions were opened up to undergraduates internationally, and the journal adopted a double-blind peer-review structure that has in recent years become an industry standard. It is a terrible misfortune that the present issue contains no discussion pieces, their being the cornerstone of the journal's purpose, but *Critique* is still the same journal as before, even if its new series is more properly organized, so it must bear the black marks on its history, which includes an archive largely unsuitable for discussion or for critique. If there is anything that I should wish of this issue, it is that it should serve as the bedrock for the new series, which, no doubt, will be replete with undergraduate discussion of undergraduate essays, and excellent reviews of professional books.

With the internationalization of the journal in its new series, dozens of submissions were made to the journal from undergraduates in Europe, Asia and North America. The manuscripts submitted for this issue were reviewed by almost two dozen reviewers from the same three continents. And those worthy of publication were, in the end, largely British and American. Such a distribution is quite unsurprising, with Europe, America, Canada, Australia and Eastern Asia representing academic bastions, behind which Southeastern Asia, the Middle East, Africa and South America have traditionally fallen behind by some significant degree. A special effort was made to encourage undergraduates from these regions to submit to the journal, but it is regrettable that the pressed publication schedule for this issue left students with little over a month to submit to *Critique*, and no doubt the number of submissions would have been much greater otherwise. However, submissions are accepted all year round, and those submissions that were too late to be published in this issue are already under consideration for the next issue, forthcoming in the winter of 2022. A call for papers will be issued several months in advance of publication of that issue, and of other issues in the future, and it is hoped that the journal will grow not only in the number of articles it publishes, but in the breadth of regions that it covers. If *Critique* is to properly serve as an antenna for undergraduate philosophy around the globe, it must first reach that far.

B.V.E. HYDE

The Pantheist Fluctuating Maximal God

CATHERINE REDFERN
Durham University

In recent literature, there has been considerable discussion concerning alternative conceptions of God. One of these is the pantheist conception of God, which has not always been considered a genuine alternative to classical theism. Generally, pantheism claims that God is identical with the universe,[1] though there are many varieties of pantheism. One such variation is that of 'agentive cosmopscyhism', as developed by Philip Goff. The essential claim of agentive cosmopsychism is that the universe designed and regulates itself via an ability to recognise and respond to reasons.[2] Another alternative conception God discussed in recent publications is Yujin Nagasawa's 'maximal God thesis', which is a form of perfect being theism. Nagasawa's thesis retains the omni-God properties of the classical conception of God, but states that they are not intrinsically maximal. Rather, God possesses the maximal *consistent* combination of omnipotence, omniscience and omnibenevolence.[3] While radical and innovative in itself, Nagasawa's thesis has been developed by Anne Jeffrey, Asha Lancaster-Thomas, and Matyas Moravec, who argue that adopting a fluctuating model of the maximal God allows the thesis greater defensibility.[4]

In this essay, I aim to join these views together by arguing that the conception of God assumed within agentive cosmopsychism is compatible with that of the fluctuating maximal God. That is, there could plausibly be a pantheist maximal God of the fluctuating variety. Firstly, I will discuss agentive cosmopsychism and the maximal God thesis in more detail and demonstrate the way in which the two are compatible. I will then explain the motivation for this proposal, namely that supplementing the pantheist position with the maximal God thesis strengthens it against significant criticisms, most notably the claim that the pantheist God is not worth worshipping. Lastly, I will consider some potential objections to these arguments as well as some possible responses. As such, I hope to illustrate the viability of the pantheist fluctuating maximal God as an alternative to classical theism.

1. Agentive cosmopsychism and maximal God varieties

In a recent paper,[5] Phillip Goff advocated a pantheist position termed *agentive cosmopsychism*, as developed from *constitutive cosmopsychism*, which holds that fundamental categorical properties are instantiated by the universe as a whole, and are consciousness-

[1] Mander, 2000: 199
[2] Goff, 2019: 108
[3] Nagasawa, 2017: 92
[4] Jeffrey et al., 2020: 232
[5] Goff, 2019

involving.[6] That is to say, consciousness is a fundamental feature of the universe. Agentive cosmopsychism retains this claim, though it is more specific in that it claims that the universe is a conscious subject and agent. This does not entail that the universe has complex mental states like human beings, rather this claim asserts that the universe has the capacity to recognise and respond to facts about value.[7] The essential claim of Goff's paper is that the universe designed itself by means of this mechanism of recognition and response, though this mechanism is limited by the laws of nature. That is, facts about the world which the universe can recognise and respond to do not include events and actions prohibited by the laws of nature.[8] As such, this picture of the universe provides us with a notion of a limited intelligent designer, though only limited in the sense that this designer cannot do anything that is physically impossible. If one accepts the pantheist agenda, this conception of the universe doubles as a conception of God.

Another contemporary conception of God is that of Yujin Nagasawa's 'maximal God.' In his book *Maximal God: A New Defence Of Perfect Being Theism,*[9] Nagasawa retains the traditional theistic claim that God is the greatest, most perfect being conceivable. However, he rejects the traditional view that God possesses inherently maximal Omni-God properties. That is, for Nagasawa, God possesses the maximal *consistent* set of omniscience, omnipotence and omnibenevolence.[10] This thesis is motivated by the contradictions that arise from analysis of omni-God properties, which can be divided into three distinct groups. Type A arguments are those which demonstrate the internal inconsistency of one of the omni-God properties. An example of a type A argument is the paradox of the stone, which demonstrates the internal inconsistency of omnipotence.[11] Type B are those which establish mutual inconsistency between two or more of the omni-God properties. For example, the so called 'argument from God's inability to sin', which maintains that omnipotence and omnibenevolence are mutually incompatible. If God is omnibenevolent, it follows that God cannot sin. Yet, if God cannot sin, God cannot be omnipotent.[12] Finally, type C are those which argue that the omni-God properties are inconsistent with a specific fact about the world.[13] The most pressing example of a type C argument is the problem of evil, which demonstrates the inconsistency of omni-God properties with the existence of evil in the world.[14] These are serious concerns for perfect being theism, and, as Nagasawa argues, they can be avoided by adopting the maximal God thesis. By definition, the maximal God is not *all* powerful, knowledgeable and loving, but *consistently* powerful, knowledgeable and loving, so inconsistency issues do not arise.

Despite being progressive and fairly successful at overcoming criticisms of perfect being theism, Nagasawa's thesis is by no means infallible. Anne Jeffrey, Asha Lancaster-Thomas and Matyas Moravec have therefore proposed an updated version, which they feel strengthens the maximal God thesis against opposing arguments. In "Fluctuating Maximal God",[15] they reject Nagasawa's picture of the maximal God as 'static' in regard to omni-God

[6] Ibid. 104
[7] Goff, 2019: 108
[8] Ibid. 109-110
[9] Nagasawa, 2017
[10] Ibid. 92
[11] Ibid. 83
[12] Ibid. 84
[13] Ibid. 91
[14] Ibid. 85-86
[15] Jeffrey et al., 2020

properties. They argue that the maximal God thesis would be more persuasive if we allowed the great-making properties to fluctuate in the course of time, as opposed to there being only one 'consistent' combination of these properties, which remains constant over time. On this view, the values of each individual property may increase or decrease in different instances, but the aggregate of the three always remains the highest possible value.[16] As such, God is still the most perfect being conceivable, though not inherently maximal. The main benefit of adopting a fluctuating maximal God rather than a static maximal God, is that it more persuasive against type-C arguments. The authors use the problem of evil to demonstrate this. On Nagasawa's original thesis, it seems that to make God's existence compatible with the existence of evil in the world, God's power must be reduced to such a great extent that God is rendered extremely weak.[17] However, if we allow God's power to fluctuate, this accommodates facts concerning evil in the world without eternally reducing God's power. For example, if at T1, there is great evil in the world, such a fact can be explained by reducing God's power to the extent that God cannot prevent such evil. Moreover, if at T2, there is less evil in the world, God's power may increase, and this increase is offset by a decrease in another omni-property. Therefore, this approach allows God to remain the greatest possible being yet allows God to be consistent with practically any fact about the world, including the existence of evil, without irreversibly reducing any of God's properties.

At this point, I must also clarify that I am not going to defend these positions in this paper. These hypotheses have been coherently defended elsewhere in recent literature, as such, my picture of the pantheist fluctuating maximal God works on the presupposition that both are independently plausible.

2. THE PANTHEIST FLUCTUATING MAXIMAL GOD

I argue that agentive cosmopsychism and the fluctuating maximal God thesis can be merged. That is, the two are consistent, and can be combined to create a plausible new conception of God which retains the essential features of both theories. Recall that agentive cosmopsychism holds that the universe/God regulates itself via a process of recognition and response to facts about value.[18] Akin to this procedure, the fluctuating maximal God thesis states that the value of the great-making properties possessed by God are caused to fluctuate by physical facts. This is well expressed by the authors, who state 'God may have a constant disposition to respond in appropriate ways to certain events such that the occurrence of those events activates the disposition'.[19] Assuming, as I have, that agentive cosmopsychism is plausible, there does not seem to be any reason to suppose that the God of agentive cosmopsychism could not recognise and respond to the types of physical facts which cause fluctuations in great-making properties. Given that the God of agentive cosmopsychism is identical with the universe, physical facts about the universe are themselves comprised by God. Furthermore, it seems plausible that such an entity would be able to recognise and respond to facts about itself, given that most physical agents possessing complex consciousness are able to do this, including human beings. For example, when a person is hungry, they recognise this physical fact about themselves, and respond to it by going to the kitchen to retrieve something

[16] Jeffrey et al., 2020: 235
[17] Ibid. 237
[18] Goff, 2019: 108
[19] Jeffrey et al., 2020: 235

to eat. Of course, the actual process of recognition and response may differ greatly between human beings and God. That said, the very fact that people possess this ability allows us to reasonably assume that the God of agentive cosmopsychism may also possess this ability, for it is also a conscious, yet entirely physical entity. If this is the case (and there is little reason to doubt that it is), agentive cosmopsychism is compatible with God reacting to any fact or physical event consistent with the laws of nature.[20] As such, so long as those conditions which cause omni-God fluctuations are physically possible (and there is no indication in the thesis that they are not), they are obviously within the scope of recognition of the agentive cosmopsychist God, for they are facts which this God itself constitutes.

Furthermore, agentive cosmopsychism is compatible with the fluctuating maximal God in that both of these views postulate a God that is somewhat limited, in the sense that they are dependent on the physical world. The God of agentive cosmopsychism is constrained by the laws of nature, where the fluctuating maximal God is constrained by the consistency of great-making properties and fluctuating facts about the world. That is, the God of agentive cosmopsychism can only act in response to a physical fact, where the value of the great-making properties possessed by the fluctuating maximal God is literally determined by the way the world is. Therefore, both rely on the world in some way. This is not the case for the God classical theism, who may act freely at any time, and who is not reliant on any other entity for their properties. As such, the fusion of the two is not contradictory for neither of these Gods are entirely self-sustaining. Therefore, I postulate the pantheist fluctuating maximal God (PFMG). This conception holds that God and the universe are identical, and that the universe is powerful, knowledgeable and benevolent to the extent that these properties do not raise contradictions. As this is a fluctuating maximal God, these properties can alter, when necessary, in response to physical facts that the universe recognises about itself.

It may reasonably be asked, at this point, why this conception adopts the fluctuating maximal God model as opposed to Nagasawa's original version. The first reason is that, as noted by the authors of "Fluctuating Maximal God",[21] the fluctuating model has more argumentative power than Nagasawa's original thesis, so it is reasonable for the panpsychist to adopt this model to avoid criticisms of the original. Furthermore, agentive cosmopsychism is more compatible with the fluctuating model than the original model, for both operate via a process of recognition and response, where the original maximal God thesis does not. Therefore, a pantheist fluctuating maximal God is more tenable than simply a pantheist maximal God both in terms of strength and coherence of the combination.

That said, some may object to the compatibility of pantheism and the fluctuating maximal God thesis. The fluctuating maximal God is a God of perfect being theism, which is distinct from the universe. The God of pantheism, on the other hand, is identical with the universe. It could reasonably be argued that whatever properties these entities possess, they are in principle incompatible. A being surely cannot be simultaneously distinct from and identical to the universe.

While this is true, there are two ways of responding to such an objection. Firstly, the defender of the PFMG merely uses the methodology taken over from the fluctuating maximal God to strengthen the position of the pantheist God. Namely, to argue that the pantheist God is worthy of worship (as will be explained in section four). If we can feasibly worship the fluctuating maximal God because it possesses omni-properties, surely, we are able to worship

[20] Goff, 2019: 109-110
[21] Jeffrey et al., 2020

anything that possesses such properties. The PFMG possesses these properties, and as such it is worthy of worship. The second possible line of defence is to deny that the fluctuating maximal God needs to be distinct from the universe in the first place. Being distinct from the universe was clearly a feature of perfect being theism, but so was timelessness. The fluctuating maximal God showed that timelessness can be easily replaced with existence in time at no cost. If anything, removing timelessness is an improvement.[22] It seems plausible that getting rid of the claim that God is distinct from the universe is also an improvement.

3. MOTIVATION

Although I have shown that the two are *compatible*, one might reasonably ask whether their combination is *desirable*. What advantage is borne by postulating such an entity as the pantheist fluctuating maximal God?

The pantheist position is argumentatively strengthened once supplemented with the fluctuating maximal God thesis. That is, the postulation of the PFMG increases the defensibility of pantheism in general, for it allows pantheists to overcome certain substantial criticisms more easily. While this part of the discussion will mainly be focused on the implications of the PFMG for agentive cosmopsychism, it is worth mentioning that this proposal also benefits the maximal God thesis. The claim that the maximal God is (or could be) a pantheist God endows the thesis with a greater degree of qualitative parsimony, for the only substance it requires is the physical universe. On this view, God is not independent of the physical world. As a result, adherents of the maximal God thesis need not postulate any extraneous, non-physical substances, as is the case in most omni-God conceptions. Of course, adherents of the maximal God thesis may see this as too great a move away from the perfect being theism that originally motivated the thesis. That said, it is still worth acknowledging that this option exists.

While the increase of qualitative parsimony in the maximal God thesis is certainly a positive side-effect of my proposal, the primary motivation for adopting this conception is inextricably linked to its utility for pantheism. Like most philosophical theories, the strength of pantheism is heavily contested. There are many substantial criticisms of pantheism discussed elsewhere in the literature, but I am going to focus on one issue in particular: the claim that pantheism is not an adequate conception of God, because the pantheist God is not worthy of worship. In his paper "The Personal Pantheist Conception of God", Peter Forrest argues that a conception of God is a description of a being 'supremely worthy of worship'.[23] This idea can also be found in the earlier work of William Rowe, who states 'central to the idea of God is that God is worthy of unreserved praise, admiration, and worship'.[24] Though it may not seem that many theists actively or knowingly endorse this definition, I hold it to be conceivable that this notion is necessarily implicit in the conception of God maintained by many theists. Forrest himself constructed from this notion his own form of pantheism. However, I believe that acceptance of this definition constitutes a substantive issue for the pantheist conception. If it can be shown that pantheism does not postulate a God worth worshipping, it follows that pantheism may well be incompatible with the instinctive values of many theists. If this is true, it would seem doubtful that pantheism offers a plausible conception of God. After all, how can

[22] Jeffrey et al., 2020: 234-235
[23] Forrest, 2016
[24] Rowe, 2005: 16

a conception of God be plausible if it is incompatible with the intuitions of those who believe in God?

This is where the PMFG comes in. It seems plausible to claim that many theists deem the classical conception of God 'worthy of worship' on the basis of the omni-God properties.[25] The word 'worship' derives from the old English word 'weorthscipe', meaning 'to ascribe worth'.[26] Worship, then, is the act of ascribing immense value to a particular entity. The qualities of power, knowledge and benevolence are deemed valuable by most human beings. It would be hard to deny, then, that an all-powerful, all-knowing, perfectly good being is worth worshipping, for it possesses the qualities we deem valuable to a higher degree than any other being. It seems obvious that many would ascribe worth to such an entity. Conversely, it is probable that those who do not regard the pantheist God as being worthy of worship take this position on the basis of the lack of omni-God properties in the pantheist conception. If not omniscient, omnipotent and omnibenevolent, what properties does the pantheist God possess which are valuable enough to ascribe immense worth to? Therefore, allowing pantheism to be compatible with great-making properties via the maximal God thesis goes a long way in overcoming this criticism. The PFMG possesses power, knowledge and benevolence, and is the greatest possible being, so would arguably be worthy of worship to traditional theists. Simultaneously, the PFMG is not susceptible to the contradictions raised by inherently maximal omni-God properties. As such, it can be plausibly argued that the PFMG meets Forrest's definition. Therefore, the PFMG proposal strengthens the pantheist position in disallowing claims that pantheism is implausible because the pantheist God is not worthy of worship.

4. Objections and responses

Let us turn to some objections against my view. Before going further, a clarificatory note is required. I accept that combining the agentive cosmopsychist position with the maximal God thesis entails that the resulting conception will be susceptible to the criticisms of both of these individual views. These criticisms are, of course, substantial. However, they are heavily discussed in the original debates concerning agentive cosmopsychism and the maximal God thesis. My purpose here is to discuss and defend the pantheist fluctuating maximal God, as such I am only going to consider in detail criticisms which apply specifically to the motivation and formulation of this conception. Furthermore, it seems possible to distinguish between two types of objection to the PFMG proposal: 1) objections to the actual formulation of this God; 2) objections to the motivation for this proposal. I will start by considering the first type of objection.

Firstly, one might object that it is not clear from the formulation of the PFMG how this type of being could possess omni-God properties. The PFMG is identical with the universe, therefore it is heavily characterised by physical properties. As far as we are aware, no physical entity has ever possessed such a high degree of knowledge or power. How, then, can such a God possess omni-God properties? Even if we accept that the God of agentive cosmopsychism (an implicit constituent of the PFMG) has the ability to recognise and respond to facts about the world which give rise to fluctuations in these properties, the question is how the universe

[25] Rowe, 2005: 16
[26] Clarensau, 2016

had these properties in the first place. Where did these seemingly non-physical properties come from? This criticism is fairly simple, but incredibly persuasive.

One could respond to this criticism, however, by simply rejecting it. This criticism represents a misunderstanding of the proposal. The PFMG does possess omni-properties, but only their highest consistent combination. This God does not possess *omni*potence, *omni*benevolence or *omni*science. Rather, it possesses power, knowledge and benevolence as far as consistency allows. As such, we do not need to explain how this God possesses omni-properties, since although it possesses omni-properties (that is, the properties that were maximal in perfect being theism), none of these properties are maximal on this proposal. It seems coherent, then, to claim that the PFMG possesses knowledge, power, and benevolence to the highest mutually consistent maxima in the same way that the God of classical theism possesses maximal omni-God properties: by definition.

An opponent may still maintain that the properties actually possessed by the PFMG require an explanation. However, such a claim would be almost entirely baseless, given that we allow the God of classical theism to possess omni-properties without adequate explanation. Of course, theologians have attempted to provide arguments in favour of God's possession of such properties, consider Anselm's ontological argument and Aquinas' argument from degrees. However, such arguments have themselves been heavily contested, and there is no unanimous agreement that they are sufficient to account for God's possession of omni-properties. Accordingly, one might avoid this criticism of the PMFG simply by offering an Anselm-style ontological argument for the PMFG. This could be formulated as follows:

1. The PMFG is, by definition, the greatest physically possible being.

2. The greatest physically possible being must possess knowledge, power and benevolence to their highest consistent maxima, otherwise it would not be the greatest physically possible being (for a being which possessed such properties would be greater).

3. Therefore, the PMFG possesses knowledge, power and benevolence to their highest consistent maxima.

While an argument of this type may spark further criticisms of its own, this does not undermine the employment of it to overcome the criticism as a whole. This argument is on par, argumentatively, with those supporting the God of classical theism, for none of them are infallible. Furthermore, this argument mirrors Anselm's defence of the God of classical theism. As such, anyone who accepts the use of the ontological argument to defend perfect being theism must accept the use of this argument to defend the PMFG. If they do not, the onus is on them to explain why.

A further objection to the formulation of the PFMG concerns the account given of the facts that the universe/God recognises and responds to. How do we know that the type of facts which cause fluctuations in omni-God properties are the same types of facts that the universe has the ability to recognise and respond to? Agentive cosmopsychism holds that God's ability to recognise and respond to reasons is constrained by the laws of nature, which implies that the type of facts God engages with are physical in essence. What if the facts about the world which give rise to fluctuations are not physical in nature, but rather are personal or emotional facts

about situations, such as 'Amy is sad'? Such facts are obviously within the scope of recognition and response of practically any non-pantheist God (including the non-pantheist maximal God theses). Yet, if it is true that the God of agentive cosmopsychism cannot respond to such facts, this is a significant problem. If this were the case, the maximal God thesis and agentive cosmopsychism would not be compatible, for it is the ability of the agentive cosmopsychist God to recognise the facts that cause fluctuations which allows the combination of the two. If these facts were not compatible, the PFMG would not be possible.

However, there is little ground for this criticism. There is no independent evidence to believe that the types of facts that spur fluctuations in omni-God properties are non-physical. Furthermore, even if there was evidence in favour of this claim, this would not necessarily undermine the PFMG. While it is true that agentive cosmopscyhism states that the laws of nature constrain the ability of God to recognise and respond to reasons, this does not necessarily entail that the only facts that God can respond to are physical ones. In fact, what it suggests is that the only types of facts that God *cannot* respond to are those which are physically impossible. It wouldn't seem coherent to argue that facts about emotions are physically impossible, for these aren't even in the physical domain. Therefore, these types of facts needn't be included in the classification of claims which the God of agentive cosmopsychism cannot recognise and respond to. Another way of countering this objection is by observing that it is likely that emotional facts will be, at least in part, physical. They may have necessarily corresponding physical counterparts or have physical effects. For example, 'Amy crying' may be a physical effect or be a physical fact necessarily corresponding to the non-physical fact that 'Amy is sad'. As such, even if it was the case that God could only respond to physical facts, God could still adequately respond to emotional facts by responding to the corresponding physical facts or physical effects.

A related objection may be raised from the possibility of the fluctuation of power. The PFMG conception adopts the fluctuating maximal God model, meaning that the degree of omni-God properties can increase or decrease as a result of situational facts. As such, God's power will decrease if the situation requires it (to allow for the dominance of another property, yet not give rise to contradiction).[27] Some may ask, then, could God's power not decrease so low that God no longer has the ability to recognise and respond to any facts at all? If this were the possible, the PFMG conception would be self-defeating. As previously mentioned, the combination of the two theses requires God to have an ability to respond to facts, yet the combination seems to entail that God could lose this ability.

A version of this criticism was posed against the original formulation of the fluctuating maximal God, though the implications are slightly different here. The response of the authors is twofold: we could either argue that God's power never decreases enough to remove the ability to respond to facts, or that God does not exert any power at all in affecting fluctuations.[28] It is possible to adapt their second response in such a way that vindicates the PMFG. Goff's original formulation of agentive cosmopsychism proposes that the universe has the ability to recognise and respond to reasons and facts, yet this original version is not an omni-God conception. This suggests that God's ability to respond to facts is independent of God's power. Since this formulation of pantheism is implicit in the PFMG conception, it can be feasibly argued that the ability of the PFMG to respond to facts that give rise to fluctuations does not depend on a certain degree of power. It therefore seems reasonable to maintain that this

[27] Jeffrey et al., 2020: 235
[28] Ibid. 245

conception is plausible, for the ability to recognise and respond to reasons and facts *allows* the pantheist God to also be a fluctuating maximal God, but the fluctuating model has no bearing on this ability.

As for objections concerning the motivation for the PFMG, there are two probable lines of argument: (i) that the PFMG does not overcome criticisms, as it is still not worthy of worship; (ii) that the PFMG is not necessary for overcoming criticisms, and as such is an extraneous proposal. In both cases, it may be argued that the postulation of the PFMG is pointless. Starting with the former, some may argue that, if the omni-God properties are not inherently maximal, the PFMG is still not worthy of worship. Traditional theists may be inclined to maintain that only a maximally omnipotent, omniscient and omnibenevolent God is worthy of worship.[29] As noted by the authors of "Fluctuating Maximal God", we can conceive of a being greater than the fluctuating maximal God, namely the God of classical theism.[30] The same holds for the PFMG. How then, can the PFMG be worthy of worship if it is not the greatest being imaginable? Moreover, if we allow such a God to be worthy of worship, does not the feature which qualifies an entity as being worship-able become arbitrary?

One could respond, however, that though the PFMG is not as great as other conceivable entities, it is worthy of worship, for it is the greatest conceivable *consistent* being. By definition, any conception of a being with greater degrees of omni-God properties at any one time than the PFMG will give rise to contradictions. Surely, then, the feature which qualifies the PFMG as worship-able is that it is the greatest being we can imagine which could exist without contradiction. If it is reasonable to worship the God of classical theism, which is inherently maximal but internally contradictory, surely it is also reasonable to worship a being which is maximal to the extent that it could coherently constitute the universe.

Conversely, even if it is true that the PFMG is worthy of worship on account of its consistency, there may still be ground for rejecting the PFMG on the basis that the existence of such an entity is not necessary. As previously discussed, the purpose of the PFMG thesis is to allow the pantheist God to be compatible with great-making properties, for many theists would only worship a powerful, knowledgeable, benevolent God. Some may argue, however, that the supplementation of pantheism with the maximal God thesis is superfluous, because the pantheist God is already compatible with such properties. For example, traditional omniscience is held by some to be compatible with pantheism. The essential claim here is that God's omniscience is indistinguishable from reality itself. This is well expressed by William Mander, who states that, if omniscience entails that God knows everything about the universe, how can we distinguish between God's knowledge and the universe that God knows? In this sense, reality *is* God's knowledge, and hence omniscience is compatible with a pantheistic world view, as all that is necessary for God to be omniscient is the universe (and this precisely what pantheism holds to exist).[31] If it is true that the pantheist God is compatible with omni-God properties independently of the maximal God thesis, it follows that the maximal God thesis is not required for pantheism to overcome the claim that the pantheist God is not worthy of worship. If this is the case, the PFMG proposal can be rejected by appeal to Ockham's razor. The postulation of the PFMG may simply overcomplicate the pantheist position, if it is not needed to overcome the issue at hand.

---

[29] Rowe, 2005: 32
[30] Jeffrey et al., 2020: 245
[31] Mander, 2000: 200

However, this criticism is unfounded. There is reason to believe that the postulation of a PFMG is not extraneous, rather, it is genuinely purposeful. Even if it is the case that the pantheist God is independently compatible with omniscience, this does not necessitate compatibility with the other omni-God properties. If God is identical with the universe, and Mander is right, it is obvious that God is all-knowing. But this does not necessarily mean God is *all*-powerful or *all*-good. Even with such a degree of knowledge, issues such as the problem of evil are still prevalent. As such, the fluctuating maximal God thesis is required to render the pantheist God consistent with power and benevolence, as well as knowledge. As such, there is sufficient motivation for postulating the combination between the maximal God and the pantheist God, for this is a plausible way of allowing the pantheist God to possess all the properties deemed worthy of worship.

4. Conclusion

In this essay, I have proposed a conception of God which combines agentive cosmopsychism with the fluctuating maximal God thesis. That is, God is a conscious agent, identical with the universe, possessing fluctuating degrees of power, knowledge, and goodness. The aggregate of these properties, however, remains at the highest consistent value. I believe that the postulation of such a God could strengthen the argument for pantheism, for it allows the pantheist God to be consistent with omni-God properties, as understood using the fluctuating maximal God thesis. As such, it allows pantheists to overcome criticisms to the effect that the pantheist God is not worthy of worship, as it is not an omni-God. I have considered some possible objections to this proposal, and in responding to these have shown that it is possible to coherently defend it. Granted, more work may be needed to solidify the strength of this conception of God. However, I believe that there is sufficient evidence to claim that the basic notion of a pantheist fluctuating maximal God represents a plausible potential development of pantheism, and a genuine alternative to classical theism.

References

Clarensau, Tyler (2016). Is God Worthy of Our Worship? Retrieved from https://tobethe.church/is-god-worthy-of-our-worship/.

Forrest, Peter (2016). The Personal Pantheist Conception of God. In Buckareff, Andrei and Nagasawa, Yujin, *Alternative Concepts of God: Essays on the Metaphysics of the Divine*. Oxford University Press.

Goff, Philip (2019). Did the Universe Design Itself? *International Journal for The Philosophy of Religion 85.1,* 99-122.

Jeffrey, Anne; Lancaster-Thomas, Asha; Moravec, Matyáš (2020). Fluctuating Maximal God. *International Journal for the Philosophy of Religion 88,* 231-247.

Mander, William (2000). Omniscience and Pantheism. *The Heythrop Journal 41.2,* 199-208.

Nagasawa, Yujin (2017). *Maximal God: A New Defence of Perfect Being Theism*. Oxford University Press.

Rowe, William (2005). Divine Power, Goodness and Knowledge. In Wainwright, William, *The Oxford Handbook of Philosophy of Religion*. Oxford University Press.

Locke's Theological Foundations[1]

J.R.H. AMERESEKERE
The University of Toronto

IN LOCKEAN SCHOLARSHIP, there subsists a substantial disagreement about the foundations of John Locke's Natural Rights and Natural Law. Locke scholars such as John Dunn, Jeremy Waldron, and others, have broadly and diversly argued that Locke's political philosophy essentially rests on (Christian) theological foundations.[2] Alternatively, scholars such as Leo Strauss, Thomas Pangle, Michael Zuckert, and others, have argued the contrary.[3] This essay strives to clarify the relationship between Locke's political philosophy and his theology, namely in the *2nd Treatise on Government*. I argue that Locke's conception of persons, their Natural Rights, and the Natural Law which protects those rights are all rooted in divine ownership of persons, and Locke is therefore a fundamentally theocentric thinker. Firstly, I will briefly account for the current consensus on the debate; secondly, I will establish Locke's unfolding political project; thirdly, I will examine it considering his relevant theological views; lastly (contrary to the historical approach of the Cambridge School or the hermeneutic approach of the Straussian school, outlined below) I will establish the inseparable link between his politics and his theology, from divine ownership and intention.

II. CURRENT CONSENSUS

Evaluating the allegedly theological foundations of Locke's political philosophy, including but not limited to the *Two Treatises*, generally operates twofold. The method has conventionally been either (i) a matter of philosophical history, evaluating Locke considering his Calvinist family-background and the social context of Calvinist traditions in England at large i.e., the approach of the Cambridge School, or alternatively (ii) evaluating Locke hermeneutically, examining the language and rhetoric he employs and the purposes they may serve i.e., the Straussian approach. The former concludes that the theocentricism of Locke's work is a natural consequence of his essentially religious socio-cultural climate in 17th-century England (a conclusion mostly compatible with my own, despite being ascertained through a different method). Dunn, for example, comments that 'there is no point in Locke's life at which

[1] I would like to express my gratitude to Erfan Xia, Henry Krahn, and Christine Park, all of whom provided substantial and always-helpful commentary on (far too) many drafts of this paper.
[2] See Dunn's *The Political Thought of John Locke* and *Locke: A Very Short Introduction*, and Waldron's *God, Locke,* and *Equality: Christian Foundations in Locke's Political Thought*.
[3] See Strauss' "Locke's Doctrine of Natural Law" and "On Locke's Doctrine of Natural Right", Pangle's *The Spirit of Modern Republicanism*, and Zuckert's *Natural Rights and New Republicanism* and his critique of Waldron's "God, Locke, and Equality".

he doubted that some men did know their duty to God',[4] suggesting the essentially theocentric *ethos* of Locke's life. The latter argues that Locke is more of a Hobbesian (or rather, 'pragmatic') proponent of natural rights and natural law merely using complex theological and religious rhetoric to disguise his unbelief or religious-indifference to avoid Christian suspicion and criticism. Stoner's appraisal of Strauss, for example, goes as far as to say that 'Strauss's account of Locke [is that of] an atheist in the mold of Hobbes and Spinoza who succeeded by his mastery of the art of esoteric writing in concealing his unbelief.'[5] I should note that the literature both between and within these schools of thought is notably extensive, and my account – while proving sufficient to explain their overarching theses – does not examine the nuances of each. Nonetheless, there is no substantial consensus between the schools of thought on the matter thus far.

While these approaches each have obvious and substantial scholarly merit, I privilege a strictly textual approach. By 'textual approach', I merely mean one that examines Locke's arguments in-themselves, independent of (i) the various socio-cultural contexts which may or may not have given rise to them, or (ii) the political or pragmatic purposes they may serve; I will simply evaluate Locke's *Treatise* by examining if his central arguments remain coherent if one omits or substitutes the overtly-theological premises present in the work. Not only does this simplify the understandable complexity of the aforementioned historical and hermeneutic approaches, but consequently requires far fewer assumptions and suppositions from cultural, intellectual, political, religious, and social climates. In simple terms, the approach I privilege will examine Locke on the fewest assumptions, thereby remaining closer to his original texts and arguments. This, in turn, will allow for a more charitable – and thereby more Lockean – analysis of his work.

III. Locke's Unfolding Political Project

Locke's *Two Treatises on Government* are one-part philosophical framework for the establishment of political organization viz. social contract, and one part a political justification of revolution. The unfolding argument of the former part, namely 'the mutual preservation of [man's] lives, liberties and estates'[6], is grounded in three essential conceptions: (i) the State of Nature, (ii) Natural Law, and (iii) Natural (and Punitive) Rights. These conceptions produce the bulk of contention surrounding Locke's allegedly theocentric foundations, and are therefore the object of my inquiry here. As a preliminary, I should note that Locke presents many of these conceptions as 'self-evident' or axiomatic; the grounds for them are far more implicit throughout the text than desirable, but I digress. This section will focus on those claims, and the proceeding section will examine those implicit grounds.

Locke begins by examining the hypothetical conception of pre-political human associations (i.e., the State of Nature), so that the conception of political organization he develops may adequately engage with the 'natural predispositions' of humans before they enter into societies. This state is categorized by two essential conditions, namely (a) perfect freedom, and (b) perfect equality. The former notion consists in allowing those in the state of nature to act (namely upon their persons and property, as Locke will establish) in accordance with their

[4] Dunn, *Locke: A Very Short Introduction*, Chapter 3.
[5] Stoner, "Was Leo Strauss Wrong About John Locke", 553.
[6] Locke, *2nd Treatise*, § 123.

own will, thereby sketching a proto-liberal notion of individual autonomy and freedom.[7] The latter refers to the symmetrical faculties, powers, and rights that individuals in the state of nature possess in relation to one another, positing a largely egalitarian conception of the individuals in Locke's hypothetical situation.[8] These transcendent, pre-political conditions implicate that each may act in accordance with their own will inasmuch as they do not infringe upon their equal, as none in the state of nature may arbitrarily exercise power over another who is an essential equal (from 'perfect equality' as defined above).[9] However, the implication is quickly made explicit.

These conditions and their implications culminate in Locke's Law of Nature – the presiding normative law 'which obliges everyone […] that no one ought to harm another in his life, health, liberty, or possessions' in the State of Nature.[10] This is the normative imperative which *actually* obliges those in the state of nature not to exercise arbitrary dominion over another. It is Locke's meta-ethical check and balance against the potential chaos that could ensue from a state of perfect freedom and equality. A crucial detail about Locke's Law of Nature is that it also follows from *what* one ought not to harm: an individual's Natural Rights viz. life, liberty, and estate. This consequently establishes Locke's essential, ontological conception of persons. In other words, Locke suggests that individuals possess the natural right to the life of their persons, the natural advantage of their liberty, and their estates and possessions – all of which are subsequently protected by the Law of Nature. Moreover, Locke presents these rights as *pre-political* in that they belong to persons even in the State of Nature, regardless of the political associations they may come into. They are, therefore, grounded in the most rudimentary natural command understood by all creatures of 'reason, which *is* that law'[11] – not some social or civic law. Thus, Locke's very first normative (ethical) commitments follow from his conception of persons.

There is one final, concerning matter for Locke's philosophy: the right of reciprocal punishment. Locke appears somewhat skeptical of how well the Law of Nature is *actually* obliged; perfect equality and freedom without a single, unified authority (i.e., the state) to enforce the normative law that regulates them is a sufficient condition for social chaos. Nonetheless, because these rights are so inviolable, it follows for Locke that failure to follow the Law of Nature (i.e., to violate another's natural rights) must be met with some reciprocal punishment at the individual level. Equal parties simply cannot justly subordinate or violate one another, and any efforts to do so are proportionally met with justified punitive action.[12] Simply put, if anyone infringes upon another, then they are subject to reactive punishment at the hand of that individual or some capable party. This right to reciprocal punishment itself has been a concern in the state of nature from its inception in Locke's *Treatises*, and is one of the fundamental motivations for the development of a social contract. The right to punish can often be abused or exercised in excess – this chaos can in turn lead to a state of war.[13] As these rights are so essential and inviolable, any such state of war and the discord it would entail would in turn infringe upon them, so Locke suggests that individuals come into voluntary associations

[7] Ibid., § 4.
[8] Ibid.
[9] Ibid.
[10] Ibid., § 6.
[11] Ibid.
[12] Ibid., § 7-8.
[13] Ibid.

for the mutual protection of their natural rights.[14] This is the essence and fundamental conclusion of Locke's project.

While the nuances of Locke's political organization are beyond the scope and interest of my work here, the grounds of his State of Nature, the Law of Nature, and Natural Rights are not, though they still remain obscure. As stated in the preliminary, Locke takes a fairly dogmatic stance to the grounds of his various philosophical assertions, starting from various axioms and 'self-evident' assertions. This, which while not essentially problematic for his purposes, leaves some clarity to be desired and requires tracing the proverbial 'fault lines' of his argument. Namely, what are the foundations of Locke's various conceptions, and are they essentially theological?

IV. Locke's Grounds for Natural Rights and Natural Law

Though having now established Locke's essential political conceptions of Natural Rights, as protected and promulgated by the Law of Nature, I will proceed to establish the justifications and foundations of those aforementioned concepts. The Lockean is therefore tasked with answering the two questions: (i) what are the metaphysical grounds for the Natural Rights viz. life, liberty, and estate, every person inviolably possesses and (ii) what promulgates the Law of Nature which protects those Natural Rights?

Locke begins with a fundamentally theistic appraisal of persons – not only as created and thereby belonging to God, but intentionally and explicitly endowed with their Natural Rights by God 'by his order and by his business'.[15] As beings created by God, persons in Locke's philosophy

> are [God's] property, whose workmanship they are, made to last during his, not one another's pleasure: and being furnished with like faculties, sharing all in one community of nature, there cannot be supposed any such subordination among us, that may authorize us to destroy one another, as if we were made for one another's uses, as the inferior ranks of creatures are for our's.[16]

Locke invokes divine ownership (and intentionality) of persons, and consequently makes the ground of his tripartite framework of life, liberty, and estate clearer. This has nuanced implications. First and foremost, persons are not their own and are therefore not entitled to destroy neither themselves nor others. Secondly, by God's intention, persons are divinely endowed with their Natural Rights (i.e., 'sharing all in one community of nature').[17] As Locke disavows the destruction of persons from a position of divine ownership, he thereby disavows any violations of the life, liberty, and estate that compose said persons. The crux of this argument is that, from the nature of persons as created in a particular way 'by [God's] order and business' with certain endowments, said endowments may never be destroyed by oneself or another inasmuch as the very persons they belong to may not be destroyed either. Given that there is a pre-political, ontological quality to these rights, belonging to a person by virtue of God's design, Locke is able to successfully tie the protection of personhood and Natural Rights

[14] Ibid. § 123.
[15] Ibid., § 6
[16] Ibid.
[17] Though Locke does not use the terms 'life, liberty, and estates' in the part of § 6. I cited, Locke's critique of God's 'order and business' exist in the broader context of those natural rights.

together in the State of Nature. However, the relationship between personhood and natural rights is by no means a loose theological connection, but clearly culminates in the very Law of Nature – a rational law of divine providence. As I have successfully shown above, Locke's normative commitments follow from his conceptions of persons, and thereby exist to protect God's property in those political persons.[18] Recall that the Law of Nature states that

> The state of nature has a law of nature to govern it, which obliges everyone: and reason which is that law, teaches all mankind who will but consult it, that being all equal and independent, no one ought to harm another in his life, health, liberty or possessions.[19]

Locke promotes a Law of Nature known by divinely-endowed faculties, which not only establishes a clear picture the Natural Rights endowed in persons pre-politically, but vehemently protects them. It is explicit in Locke's works that the 'order and business' of God – His creations – are explicitly and categorically protected by natural, divine law. This often overlooked argument from divine ownership over persons is at the core of Locke's theory of natural rights and the law which protects them.

An essential characteristic of Locke's appraisal of life, liberty, and estate that I want to emphasize is their intrinsic goodness (or value). As these rights are God-given, pre-political, and thereby pre-industrial, they exist as good by their own merit as opposed to holding value by the merit of their consequence as a 'social currency'. Therefore, it would be incorrect to state that these rights are instrumental goods, which serve some discretionary social end. They are valuable, ontological components of persons as given by the order and business of their creator, and are not to be infringed upon – an axiom known to us by the faculty of reason endowed unto us by that very God – presupposing any arbitrary social context which may follow from them. The inference is that social existence follows from these rights, not vice-versa; the very purpose of the *Treatises* is the preservation of pre-political Natural Rights, as the state of nature is not an argument for utopia but a hypothetical thought-experiment to determine the criteria of legitimate political organization. These rights exist as inviolable, suggesting that they are intrinsically valuable and promulgated and protected for their own sake. This detail provides crucial as it stands to show that the intrinsic theocentric nature of Locke's natural rights necessitate God, as I will argue in Section IV.

There is an obvious normative quality to this reading, as the intrinsic value of these rights are protected by a rational moral imperative (i.e., the Law of Nature). This informs us of two things. (i) Locke's baseline perspective of morality is a theocentric-universalist perspective, known through the God-given faculty of reason, applying in force at all times to all people regardless of any such socio-political and/or cultural context, and (ii) the law is grounded in the order, business, and/or mandate of God. *Prima facie*, Locke is a theistic thinker – a trait obviously and overtly apparent in his political writings and arguments at large. Human persons belong to God, are endowed with certain natural rights and faculties, of which are protected by natural (moral) law. However, with certain Locke scholars objecting to a necessarily-theistic reading, it is essential to examine the possibility of an atheistic reading.

---

[18] See "Locke's Theological Foundations" above.
[19] Ibid.

V. Can Locke be Understood Atheistically?

As stated in Section II, I will be answering the question: 'can Locke be understood atheistically?' by examining if his same conclusions can follow without theistic premises. Initially and most obviously, Locke cannot rely upon 'created persons' – the foundation of his framework – in the absence of some sort of creative divine being. This is the initial difficulty of charitably appraising any sort of atheistic reading of Locke's political philosophy. Locke understands the very subject of political philosophy – rational persons – as extant only by virtue of (the Christian) God. The notion of created persons and consequently divine ownership as expressed in various passages in the *Treatises* would have had to be entirely *omitted* in the case of an Atheistic framework. From here, there is no real ground for personhood, Natural Rights, and the Law of Nature which protects those rights, at least in the way Locke seems to have intended. While this may seem like such an obvious technicality, and hardly a compelling examination of a possible atheistic reading of Locke, it nonetheless has concerning implications for Lockean Atheists. It is only so obvious precisely because it is so central to Locke's writings; the 'business and order' of God is the premise against which many of his other conclusions firmly rest. Nonetheless, I will still engage with the possibility of (i) the Natural Rights of persons and (ii) the Law of Nature which protects them, arising atheistically.

Regarding the former, the right to life, liberty, and possessions are essentially tied to the apparent ontological 'order and business' of God and the Law of Nature which protects and promulgates them, and in a strictly atheistic reading, would be absent. Could they arise alternatively? An objector to my position may point out that Locke's state of nature is an essentially social, cooperative one (contra-Jean Jacques Rousseau or Thomas Hobbes) and has traditionally been read as such; they may thereby argue that Locke is best (and correctly) read as a thinker establishing natural rights in the context of an essentially social state of nature, in a way that does not *necessitate* God. They could argue that a 'social' state of nature is an adept substitute for 'divine providence', and these rights therefore do not demand God as much as they demand some arbitrary social context which could give rise to them *ipso facto*; perhaps humans naturally develop a reverence for themselves, their equal persons, their freedoms, and their possessions, as these things come to compose the societies which they enter into. This is a coherent way to synthesize the development of natural rights (as thinkers like Rousseau have shown), and are seemingly in line with Locke's State of Nature – namely its perfect freedom and equality.[20] Under this conception, God (regardless of whether he exists or not) is not 'necessary' for the development of Natural Rights, and Locke's account would remain coherent in His absence.

I reply to such two-fold. First and foremost, just because these rights are coherently understood socially does not mean they *arise from* the social framework which makes sense of them. I have shown above that these rights are strictly pre-political, ontologically belonging to agents by virtue of their particular nature as created persons. The rights that potential Rousseauvian objectors suggest are hardly pre-political, and require at least some collective context; this clearly contravenes Locke's intentions. Moreover, it is impossible to argue that Locke's conception of natural rights develops from the arbitrary contingency of social organization, as they hold an intrinsic quality as opposed to an instrumental social quality as I have argued above.[21] They strictly rely on the intrinsic 'design' or intentionality of persons

[20] Rousseau, *Discourse on Inequality* and *The Social Contract*.
[21] See "Locke's Theological Foundation" above.

created by God. This condition is simply too binding to overlook or circumvent, at least charitably. In truth, if one was to argue this way, Locke's natural rights would no longer possess any of the qualities he has ascribed to them. This reading simply does not stand, and fails to account for the nuances of Locke's central political concepts.

This objection nonetheless has a much more fundamental issue to circumvent. Even if one is to assume that these rights are actually a natural *telos* of man as social creatures and the argument from social development could be accepted as coherent, there is still disagreement on whether Locke's state of nature is *actually* social or not. J.J. Jenkins, for example, appraises Locke's description of the state of nature as noticeably ambiguous, at times discussing man as cooperative while simultaneously categorizing them as quarrelsome (thereby entirely necessitating the idea of a state itself).[22] It is apparent that Locke himself was rather skeptical of this allegedly social quality to his State of Nature, as he is weary of a possible State of War. Without a firm, undisputed social-position to rest on, the objection is entirely baseless. While it is possible to argue for Natural from an atheistic, social position, it is hardly plausible while staying true to the *Treatises*. Thereby, an atheistic reading of Locke's Natural Rights in this way is entirely too implausible.

In the same vein as Natural Rights, one must ask if Locke's Law of Nature could arise, or if it would even have a coherent place in the *Treatises*, atheistically. Recall that Locke's particular imperative that 'no one ought to harm another in his life, health, liberty or possessions'[23] is essentially tied to the ontological nature of persons as dependant on the order and business of God. As I have argued, Locke's normative commitments follow *directly* from his conception of persons – which, as shown above, are essentially theocentric. Recall that this natural law exists to serve two purposes: (i) the obvious protection of Lockean natural rights, but moreover, (ii) respect divine ownership over persons. Locke writes that individuals 'are [God's] property whose workmanship they are, made to last during his, not one another's pleasure.'[24] This is the underlying meta-ethical justification of the law of nature as so central to Locke's framework. From divine ownership of persons is immediately by '[man] has no liberty to destroy himself, or so much as any creature in his possession, yet when some nobler use than its bare possession calls for it'.[25] Neither the self nor the other belongs to oneself nor one another to destroy or dispose of in any way. The very content of Natural Law arises by the fact that our persons explicitly belong to God. The Natural Rights endowed in persons by the order and business of God imply both divine ownership over persons, and therefore produce the Law of Nature which protects the fruits of God's labour.

Beyond the fact that Locke's normative commitments are essentially contained in his conception of persons, even some atheistic alternative to Locke's law of nature could only be formulated as rule-consequentialist laws. On Lockean grounds, this is problematic. If I assume that Natural Rights (and therefore Locke's conception of persons) arose atheistically from some arbitrary social context (á la Rousseau), then the Law of Nature that protects them would only follow as a arbitrary social law serving a discretionary social end, and hardly a 'natural law' in truth. In the absence of divine ownership of persons, Natural Rights endowed in a person (if any at all) and the 'Law of Nature' that protects them would hold a far more instrumental quality. However, as I have argued, Locke's appraisal of the life, liberty, and possessions of

[22] Jenkins, "Locke and Natural Rights", 149-150.
[23] Ibid., § 6.
[24] Ibid.
[25] Ibid.

persons, and the Law of Nature which promulgate them hold a far more intrinsic value, valuable for their own sake for the preservation themselves. They belong to persons *a priori* and hold no real social-currency due to God's intentionality in people's innate design. It would be difficult to derive this sort of intrinsic value, as it pertains to Locke's particular conception of the Law of Nature, in the absence of divine ownership over persons. Therefore, while an atheistic appraisal of Locke's Law of Nature is hypothetically possible (as Rousseau, David Hume, Baruch Spinoza, etc., have similarly shown), it would largely fall short of the constraints he has set forth in the texts. Therefore, such a Law of Nature would hardly be a *Lockean* Law of Nature, even in the broadest sense.

Locke's argument is thereby essentially incoherent to read atheistically. One would have to first overlook the most substantial ground – divine ownership over created persons – for Locke's conceptions both the Law of Nature and Natural Rights. This strictly theistic reading is necessary as it is the foundational premise of Locke's unfolding argument. It is strictly the order and business of God which endows individuals with their Natural Rights to life, liberty, and estate, thereby formulating natural law in accordance with those two considerations. Could both Natural Rights and the Law of Nature arise entirely arbitrarily as a product of social organization or as an atheistic rational construct? It is obviously possible as thinkers such as Rousseau and Hobbes have shown in various cases, but it requires too many assumptions for a Lockean reading. These concepts would be entirely without foundation in the absence of Locke's theistic appraisal of persons and possible alternatives require too many unsubstantiated presumptions. My theistic reading stands, while the aforementioned objections do not.

VI. Conclusion

Prima facie, Locke's writing has an overtly theocentric quality to it; the notion of divine ownership of persons is at the very core of Locke's political (and ontological project). In light of the Cambridge School, there is no great shock that the socio-cultural climate of Locke's Protestant, 17th-century England was an influence on his thought; nonetheless, Locke's unfolding argument moves beyond being a mere product of context. It is deeply, and essentially tied to views grounded in Christian theology (regardless of why these views were held). While skepticism of the genuine roots of his theocentricism is always encouraged, to deny its involvement in Locke's works fails to trace the nuances of many of his arguments. There are no coherent substitutes or omissions that remain charitable to Locke's work, as I have shown. Simply put, Locke is an essentially theistic thinker (with most of his theology reflecting the Christian-tradition), inasmuch as those theological foundations are inseparable from his conclusions in the *Two Treatises on Government.*

References

Dunn, John. *Locke: A Very Short Introduction.* Oxford: Oxford University Press, 2003.

Dunn, John. *The Political Thought of John Locke.* Cambridge: Cambridge University Press, 1993.

Hobbes, Thomas. *Leviathan*. ed. E.M. Curley. London: Hackett Publishing, 1994.

Jenkins, John J. "Locke and Natural Rights" *Philosophy* 42, no. 160, (1967): 149-154.

Locke, John. *Second Treatise on Government*, edited by C.B. Macpherson. London: Hackett Publishing Company, 1980.

Pangle, Thomas. *The Spirit of Modern Republicanism.* Chicago: University of Chicago Press, 1988.

Waldron, Jeremy. *God, Locke, and Equality: Christian Foundations in Locke's Political Thought.* Cambridge: Cambridge University Press, 2002.

Rousseau, Jean-Jacques. *Discourse on the Origin of Inequality*, translated by Donald A. Cress. Indianapolis, Indiana: Hackett Publishing Company, Inc., 2011.

Rousseau, Jean-Jacques. *The Social Contract*, translated by Donald A. Cress. Indianapolis, Indiana: Hackett Publishing Company, Inc., 1988.

Stone, James R. "Was Leo Strauss Wrong About John Locke," *The Review of Politics* 66, no. 4, (2004): 553-563.

Strauss, Leo. "Locke Doctrine of Natural Law," *American Political Science Review* 52, no. 2, (1958): 490-501.

Strauss, Leo. "On Locke's Doctrine of Natural Right," *The Philosophical Review* 61, no. 5 (1952): 475-502.

Zuckert, Michael P. *Natural Rights and New Republicanism*. Princeton: Princeton University Press, 1994.

Zuckert, Michael P. "Locke—Religion—Equality," *The Review of Politics* 67, no. 3, (2005): 419-431.

Do We Have an Obligation to Mitigate Existential Risks?

SERENA K. H. SMART
Durham University

INTRODUCTION

Many people believe we have a 'moral reason' ($reason_m$) to mitigate existential risks because of the catastrophic consequences of an existential event occurring. If an existential event occurred, it would cause humanity to go extinct. As this is a relatively catastrophic consequence, it seems plausible that we have some $reason_m$ to mitigate existential risks to prevent humanity from going extinct. However, Derek Parfit advances that the consequences of an existential event are *even worse* than they intuitively appear, due to the many future lives that could come to exist, but won't, if humanity goes *prematurely* extinct. Consequently, he argues that we have an exceptionally strong $reason_m$, or *obligation*, to mitigate existential risks.

In the following article, I attempt to refute Parfit's argument from additional lives (AAL) to demonstrate that the consequences of an existential event are nowhere near as catastrophic as he envisages. Consequently, I suggest that our only $reason_m$ to mitigate existential risks is to stop humanity from going *extinct*. As the strength of our $reason_m$ to mitigate existential risks is a function of the badness of the consequences of an existential event, I argue that this greatly reduces our $reason_m$ to mitigate existential risks to the extent that it may no longer be considered an obligation. Finally, I question whether we have *any* $reason_m$ to mitigate existential risks, as the *ultimate* extinction of humanity is inevitable, and we cannot have a $reason_m$ to prevent the inevitable. I conclude that even if one rejects this final line of reasoning, mitigating existential risks is still a supererogatory act.

I: OUR SUPPOSED OBLIGATION TO MITIGATE EXISTENTIAL RISKS

An existential risk is a risk that threatens to 'annihilate Earth-originating intelligent life'. [1] Many argue that we have a duty to mitigate existential risks because of the devastating consequences of such risks occurring.

The argument reasons as follows: [2]

P1: The probability of an existential risk occurring is exceptionally high

P2: The consequences of an existential event would be exceptionally bad

[1] Bostrom, 2001, p. 2.
[2] Bostrom, 2013.

P3: We have a very strong $reason_m$ to try and prevent exceptionally bad, highly probable, things from happening

A: If a $reason_m$ is strong enough, it constitutes an obligation

C: We have an obligation to mitigate existential risks

The high number of risks in contemporary society — from nuclear war to an uninhabitable earth due to climate change — means that the probability of an existential event occurring within the next few hundred years is exceptionally high **(p1);** roughly '1 in 6.'[3] If an existential event occurred, it would cause mankind to go extinct; this consequence seems exceptionally bad **(p2).** If we have a very strong $reason_m$ to prevent very terrible, highly probable, things from happening **(p3)** — assuming that a very strong $reason_m$ is essentially an obligation **(a)** — it follows that we have an obligation to mitigate existential risks **(C)**.

Thus, our $reason_m$ to mitigate existential risks is a function of the severity of the *consequences* of an existential event; and the *likelihood* of an existential event occurring. If the probability of an existential event occurring is held constant, this gives rise to the following premise:

Pd: *The more catastrophic/worse the consequences of an existential event, the greater our $reason_m$ to mitigate existential risks.*

Although this idea is not explicitly elucidated by Nick Bostrom, it seems like a logical extension of his argument. For example, if I know that when I walk to the shops there is a high probability that I will be struck by lightning and die, I have a very strong reason to mitigate this risk by not going outside. Conversely, if the consequences of me going to the shops are that it's highly probable I will step in a puddle, my incentive to mitigate that risk is greatly reduced. I do not deem the consequences as catastrophic, and so my reason to mitigate the risk is reduced. Assuming my prudential 'reason' is comparable to humanity's $reason_m$, this analogy illustrates that our $reason_m$ for mitigating existential risks is determined by the badness of the consequences of an existential event.

II: How Bad Would the Consequences of an Existential Event Be?

Section I illustrated that the strength of our $reason_m$ to mitigate existential risks is a function of the badness of the consequences of an existential event. The following section introduces the two principal arguments outlining how bad an existential event would be and considers how strong our $reason_m$ to mitigate existential risks is in light of these arguments.

II.A) Bad

As the extinction of humanity is a pretty bad consequence, it appears we have a notable $reason_m$ to mitigate existential risks; this conclusion is intuitive.

[3] Ord, 2020.

Consider two situations, A and B. In A, one innocent person dies. In B, ten innocent people die. Assigning each person's life a value of 1, the disvalue caused by their deaths generates the following:

		Value
A:	-	1
B:	-	10

According to Utilitarianism, the best outcome is the one that maximises utility, or value. Therefore, **B** is worse than **A**, as it has a lower overall value. This supports our intuition that the more people who die, the worse an outcome is. As the extinction of humanity would cause 7.8 billion people to die, the extinction of humanity seems very bad. Following the logic outlined in **Pd** gives rise to my **intuitive reason$_m$:**

> *'We have a notable reason$_m$ to mitigate existential risks because of the significant disvalue produced by all the existing people/population who would die if an existential event occurred.'*

II.B) FROM BAD TO WORSE

Intuitively, the extinction of humanity is bad because it causes a great number of existing people to die. However, Derek Parfit argues that the extinction of humanity is *even worse* than it initially appears because of the disvalue produced by all the lives that *would have come to exist* but don't due to the *premature* extinction of humanity.

As Parfit articulates:

> The Earth will remain habitable for at least another billion years. Civilisation began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilised human history. The difference between [99% of the population dying] and [100% of the population dying] may thus be the difference between this tiny fraction and all of the rest of this history.[4]

If Parfit is correct — and the number of existing/past lives, (P), are *dwarfed* by the number of all possible future lives, (fP) — the amount of disvalue generated by the premature extinction of humanity does seem far greater than our intuitions suggest.

Expressed numerically:

	Value
1: No one dies	0
2: 99% of the population dies	-0.99(P)
3: 100% of the population dies	-1(P) + (-1(fP))

Once we account for the disvalue created by all the future possible people that could have come to exist, but don't, due to the premature extinction of humanity; the difference in disvalue between 2 and 3 is far greater than the difference between 1 and 2. Consequently, Parfit concludes that the extinction of mankind within the next hundred years is uniquely

[4] Parfit, 1984, p. 454.

catastrophic because it is *premature:* it means that most future lives will never come to exist. Not only does humanity's extinction produce a huge amount of disvalue from the many existing people who are killed, but it also produces an even greater amount of *additional* disvalue due to the loss in all potential *future lives.* An outcome that produces this enormity of disvalue is exceptionally bad; far worse than the initially conceived outcome of 7.8 billion people dying.

Applying **Pd,** it follows that our $reason_m$ to mitigate existential risks is stronger than it intuitively appears. This gives rise to my **strong $reason_m$:**

> ***'**We have a very strong $reason_m$, and thus, **obligation** to mitigate existential risks because of the huge amount of disvalue produced by all the existing people/population who would die if an existential event occurred **combined with** the additional disvalue produced by all the future lives that will never come to exist if humanity goes prematurely extinct.'*

III. Questioning the Value of Future Lives

If the **strong $reason_m$** presented in ***II.b*** is true, we have an obligation to mitigate existential risks. However, the **strong $reason_m$** rests on the claim that there is value in the creation of potential future lives and, consequently, disvalue in failing to create those lives.

The following section introduces the Intuition of Neutrality (IoN) that challenges this claim. Ultimately, if the IoN is shown to be correct, it undermines the AAL as it proves that potential future lives do not have value.

III.A) Do Future Lives Have Value?

When arguing that the extinction of humanity would produce incredible amounts of disvalue, Parfit assumes that there is value in adding more lives to the world and, consequently, disvalue in failing to add those lives. This assumption seems misguided. Surely, 'adding a person to the world is not valuable in itself, even if the person would enjoy a good life.'[5]

This is the sentiment articulated by John Broome's 'IoN', which holds that 'adding a person to the world is very often ethically neutral.'[6] As Broome clarifies, this entails that 'a world that contains an extra person is neither better nor worse than a world that does not contain her but is the same in other respects.'[7]

The IoN is highly plausible, as it coheres with our intuitions regarding procreation asymmetry: whilst we should refrain from creating individuals who would lead a very low quality of life, there is nothing morally reprehensible about choosing not to create a good life.[8] Indeed, if this principle is incorrect — and adding more positive lives to the world *is* intrinsically good — it follows that *failing* to add more positive lives to the world *is intrinsically bad.* However, this generates the absurd conclusion that a woman who chooses not to have children is doing something bad — *impermissible* even — and is thus morally blameworthy. Few would defend this conclusion, as it seems clear that the woman is violating no moral

[5] Broome, 1984, p. 167.
[6] Broome, 2004, p. 142.
[7] Ibid, p. 401.
[8] Broome, 1994, p. 167.

principle by exercising her autonomy and not having children. As this reductio ad absurdum illustrates, the idea of procreation asymmetry grounded in the IoN is highly plausible.

Moreover, as the IoN applies solely to the *coming into existence* of a person 'setting aside its effects on other people',[9] it does not commit us to the equally implausible conclusion that a person's coming into existence is *in no way* valuable. Consider a couple, *Brangelina*, who want to have a child. In world A, they manage to have a child and are exceptionally happy because of their new baby. In world B, they are unable to conceive and live a comparatively miserable life as they never fulfil their dream of becoming parents. It seems obvious that world A is better than world B, as the couple is far happier when they have a child.

Some may dispute that the IoN contradicts this conclusion, as it states that a world that contains an extra person is 'neither better nor worse than a world that does not contain her'.[10] However, we can explain why A is better than B with reference to the fact that the welfare of *Brangelina* — an already-existing couple — increases *due to* the existence of their child. This does nothing to compromise the validity of the IoN, which merely states that there is 'no consideration stemming from the wellbeing of the person herself that counts either for or against bringing her into existence'.[11] We can maintain that the coming into existence of the child is *intrinsically* neutral, whilst also claiming that the child's existence is valuable due to the positive *instrumental* value they create for existing people. Similarly, as the IoN focuses on the *addition* of individuals, it does nothing to undermine our commitment to the wellbeing of *existing* people. Thus, we can maintain that 'people are valuable, but creating them is not'.[12]

Evidently, the IoN holds strong intuitive appeal, as it coheres with our intuitions regarding procreation asymmetry; and allows for our intuition that, whilst creating life is ethically neutral, adding people to the world may still be valuable in so far as it produces instrumental value for currently existing people.

III.B) Broome's Objection from Transitivity

Despite its initial plausibility, Broome contends that the IoN is false as it violates the principle of transitivity. He argues that if a 'person's existence is neutral… the value of [a change that causes a new person to exist] must be given by its value to existing people'.[13] From this, Broome derives the constituency principle: a state of affairs is equal/better to another state of affairs only if the other state of affairs is equal/better for the people who exist in both.

Consider example **'Y',** showing 3 possible worlds:

$$1A\text{: } (w_1, w_2, \ldots, w_n, \Omega)$$
$$1B\text{: } (w_1, w_2, \ldots, w_n, 1)$$
$$1C\text{: } (w_1, w_2, \ldots, w_n, 2)$$

Within the vectors, 'w1,…wn' represents all the people who exist at the point of comparison of the three worlds; corresponding places in the vectors denote the same people. Broome asks us to consider adding an extra person, call her Grace, to one of the populations.

[9] Broome, 2004, p. 144.
[10] Broome, 2005, p. 401.
[11] Broome, 2004, p. 114.
[12] Broome and Morton, 1994, p. 197.
[13] Broome, 1984, p. 168.

In 1A, Grace does not exist and is denoted by Ω. In 1B and 1C, Grace does exist, and her welfare level is denoted by 1 and 2 respectively.

Which world is better?

According to the constituency principle, we can only compare two states of affairs by considering the people who *exist in both.* As Grace exists in 1B, but not in 1A, we can only compare 1A and 1B by considering the welfare of the people who exist in both: 'w1…wn'. Their welfare is exactly comparable, meaning 1A is equally as good as 1B. By the same logic, 1A is equally as good as 1C. As Grace *does* exist in both 1B and 1C, we can take her welfare into account when comparing these outcomes. As 'the goodness of an alternative depends only on the good of the people who exist in that alternative',[14] and Grace has a higher level of welfare in 1C, it follows that 1C is better than 1B.

However, this conclusion cannot be right, as it violates the principle of transitivity. If 1A=1B, and 1A=1C, 1B must be equal to 1C. However, the constituency principle suggests that 1C > 1B; this is a logical contradiction. As Broome summarises 'as a matter of logic, the relation 'equally as good as' is transitive, and the constituency principle implies it is not. Therefore, the constituency principle is false'.[15] As the constituency principle is derived from the IoN, the IoN must also be false.

III.C) Defending the IoN

We have two options to try and defend the IoN from Broome's objection from transitivity. Firstly, we can reject Broome's 'equal goodness' interpretation of the constituency principle, maintaining instead that two states of affairs can be *incommensurate.*[16] Alternatively, we can reject the assertion that the betterness relation is necessarily transitive, suggesting instead that comparisons between the goodness of two outcomes can be context-dependent.[17] Section **IV** expands on the argument from incommensurability, whilst Section **V** considers whether we can reject the transitivity of the betterness relation.

IV) Incommensurability and the Intuition of Neutrality

The following section defines incommensurability and illustrates how an incommensurate interpretation of the IoN (IoN.i) can overcome Broome's objection from transitivity. Next, it considers two further objections against the IoN.i: the objection from greedy neutrality and incommensurability.

IV.A) Defining Incommensurability

Two things are incommensurate if 'neither is better than the other, yet they are also not exactly equally good'.[18] Generally, incommensurability arises due to a lack of common measure, whereby the values — or bearers of value — being compared are so distinct that 'an

[14] Broome, 1991, p. 167.
[15] Broome, 1994, p. 170.
[16] Rabinowicz, 2009.
[17] Temkin, 1987.
[18] Broome, 2005, p. 407.

ordinal comparison or ranking is [not] possible'.[19] Thus, two situations can be incommensurate if they lack a single common measure of comparison or involve 'values with [distinct] qualitative dimensions that give rise to incomparability.'[20]

The notion of incommensurability is very credible, as we often feel we cannot compare things when the things being compared are suitably distinct. Consider the question: 'who is the better artist, Picasso or Mozart?'[21] This question seems impossible to answer because the artwork of Picasso (painting) and the artwork of Mozart (music) are completely different. One cannot conclude that one artist is better than the other, or that they are equally good, because their artwork lacks a common measure of comparison. As Johan Frick articulates, the category of 'painter' and 'musician' are incommensurate as the domains are 'sufficiently different in nature that a precise comparison between their goodness seems impossible'.[22] As the category of 'musician' and 'painter' are suitably distinct, one can no more conclude 'Mozart is a better artist than Picasso' than one can say 'red is better than loud.'

Ultimately, two states of affairs are incommensurate if they are neither better than, nor worse than, nor equally as good as each other. When two things are incommensurate, it essentially means that an exact comparison or ranking between them is not possible; the current orthodoxy suggests that this incommensurability can arise for three reasons. Firstly, two states of affairs can be incommensurate because the states of affairs contain values that are themselves incommensurate. For example, is not possible to perfectly compare a political system that pioneers liberty with one that champions equality, because the values of 'liberty' and 'equality' are themselves incommensurate. Secondly, two states of affairs can be incommensurate if they are suitably distinct in nature; this is the rationale motivating the incommensurability in the Mozart/Picasso example. Finally, two states of affairs can be incommensurate if there is no single point of comparison that combines all the relevant considerations between the two states of affairs. For example, when asking 'who is the better footballer, x or y?', multiple considerations are pertinent for informing the answer; whilst one player may be a great attacker, the other could be great at defence. As there is no single point of comparison that can include all relevant considerations, we could respond that the two footballers are *incommensurate.* Ultimately, if any of the three aforementioned conditions obtain, we can say that the two states of affairs are incommensurate.

IV.B) Revisiting Broome's Objection from Transitivity

Broome's objection from transitivity states that the IoN is not tenable because it violates the principle of transitivity. This argument relies on an equal goodness interpretation of the IoN, that proposes for 'two distributions that have the same population, if one of them is better than the other for someone [Grace], and at least as good as the other for everyone [w1…wn], then it is better'.[23] As example Y in section ***III.b*** illustrated, this results in a logical contradiction.

However, we can avoid this contradiction if we maintain that the two distributions are *incommensurate*. As ***IV.a*** established, when two states of affairs are incommensurate, it means

[19] Hseih, 2016.
[20] Chang, 1997, pp. 16-17.
[21] See Frick, 2017, p. 407.
[22] Frick, 2017, p. 14.
[23] Broome, 2004, p. 58.

that 'they are not equally good and neither is better than the other'.[24] Thus, we can generate the following incommensurate interpretation of the IoN (IoN.i):

> '*Ceteris paribus, the world with added people at wellbeing levels within the neutral range*[25] *is incommensurate with the world not containing these people.*'[26]

Thus, rather than arguing that 1B and 1C are *equally as good as* 1A; one can just argue that they are incommensurate with one another, as 1B and 1C contain people that do not exist in 1A. This saves us from the logical contradiction outlined in ***III.b***: if 1A is incommensurate with 1B and 1C, then we can happily conclude that 1C > 1B without violating any principles of transitivity.

Nevertheless, Broome contends that the IoN.i is ultimately untenable for two reasons. Firstly, Broome argues that the IoN.i gives rise to a kind of 'greedy neutrality' that has implausible implications for the way we value populations. Secondly, Broome questions whether it is even *possible* to have an incommensurate interpretation of the IoN, as it contradicts his assumption that the betterness relation is complete. The following subsections attempt to defend the IoN.i from these final two objections.

IV.C) Greedy Neutrality

Broome argues that we should reject the IoN.i because it gives rise to a kind of 'greedy neutrality'. His argument[27] reasons as follows:

Consider the following three distributions:

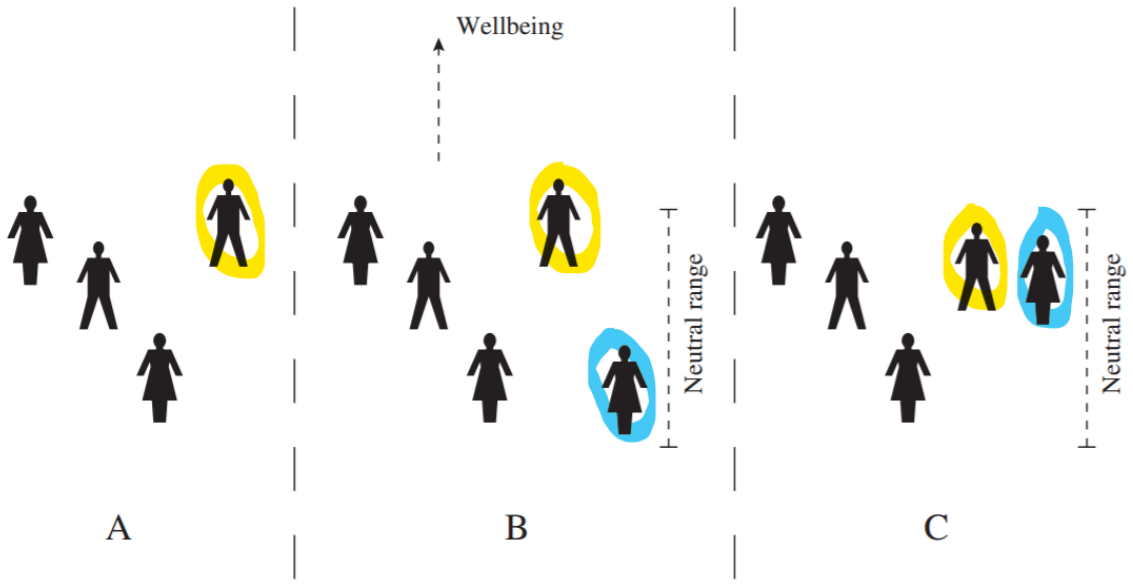

Figure 1: Three worlds depicting existing (yellow) and additional (blue) individuals with differing levels of welfare

[24] Rabinowicz, 2009, p. 4.
[25] Broome's original formulation of the IoN applies only to some narrow range of wellbeing levels, in which 'life at this level is neither better nor worse than [the person] not living at all' (Broome, 2004, p. 142).
[26] Rabinowicz, 2009, p. 4.
[27] Broome, 2005, pp. 407-409.

B and **C** contain a woman, call her *Blue,* that does not exist in **A**. Broome advances that applying the IoN.i gives rise to statement (1):

***(1) C** is neither better than, nor worse than, **A**. It is also not equally as good.*

Now compare **B** and **C.** In **B** the man, call him *yellow,* has higher welfare than he has in **C.** Conversely, *blue* has lower welfare in **B** than she does in **C**. Broome asks us to assume that, in moving from **B** to **C,** *blue's* increase in welfare more than makes up for *yellow's* small loss in welfare. On this assumption, we can derive another statement:

(2) C** is better than **B

Finally, compare **A** and **C.** As the diagram illustrates, *Yellow* is worse off in **C** than he is in **A.** Thus:

(3) C** is, in one respect, worse than **A

However, as **C** contains a person, *Blue,* who does not exist in **A**, **C** and **A** must *ultimately* be incommensurate. Therefore, we must conclude, that **C** is not worse than **A (1)** even though **C** is, in one respect, worse than **A** (**3**).

Broome argues that this conclusion is highly unintuitive. If 'C is worse than A in one respect, and neither better nor worse in the other respect; intuitively, C must on balance be worse than A'.[28] This follows from Broome's assumption that 'a bad thing plus a neutral thing must add up to a bad thing.'[29] As the IoN.i commits us to concluding that A is not worse than C, Broome concludes that it is needlessly *greedy:*

> We have found that our neutrality is greedy. Although neutral in itself, it is able to swallow up bad things and neutralize them...a bad thing plus a neutral thing must add up to a bad thing…however [the IoN.i forces us] to conclude the opposite.[30]

Thus, Broome concludes that we should reject the IoN.i as its 'greediness' has implausible implications for the way we value population.

IV.D) NOT SO GREEDY

Broome argues that we should reject the IoN.i because it is unacceptably greedy. I contend that Broome overstates the IoN.i when developing this argument; interpreted correctly, the IoN.i does not commit us to any unacceptable conclusions regarding greediness.

Broome asserts that we should reject the IoN.i because it commits us to the claim that C is not worse than A when, intuitively, we want to conclude that 'C is on balance worse than A.'[31] Thus, the argument from greedy neutrality rests on the following 2 premises:

---

[28] Ibid. p. 409.
[29] Ibid.
[30] Ibid.
[31] Ibid.

(1) The IoN.i commits us to saying that C is incommensurate with A, meaning C is neither better than, nor worse than, nor equally as good as A

(2) C is *overall* worse than A, as a bad thing plus a neutral thing must equal a bad thing

Whilst premise (2) seems plausible, I believe that the argument from greedy neutrality ultimately fails because premise (1) is false.

Recall Broome's original formulation of the IoN.i:

'A world that contains an extra person is neither better, nor worse than a world that does not contain her but is the same in other respects'[32]

The qualification 'but is the same in other respects' establishes an important constraint on when the IoN.i commits us to saying that two worlds are incommensurate. Specifically, it suggests we can only say that two worlds are incommensurate when the following 2 conditions obtain:

a) One world contains an extra person that the other world does not

b) The worlds are the same in other respects

Consequently, let us return to Broome's example of greedy neutrality:

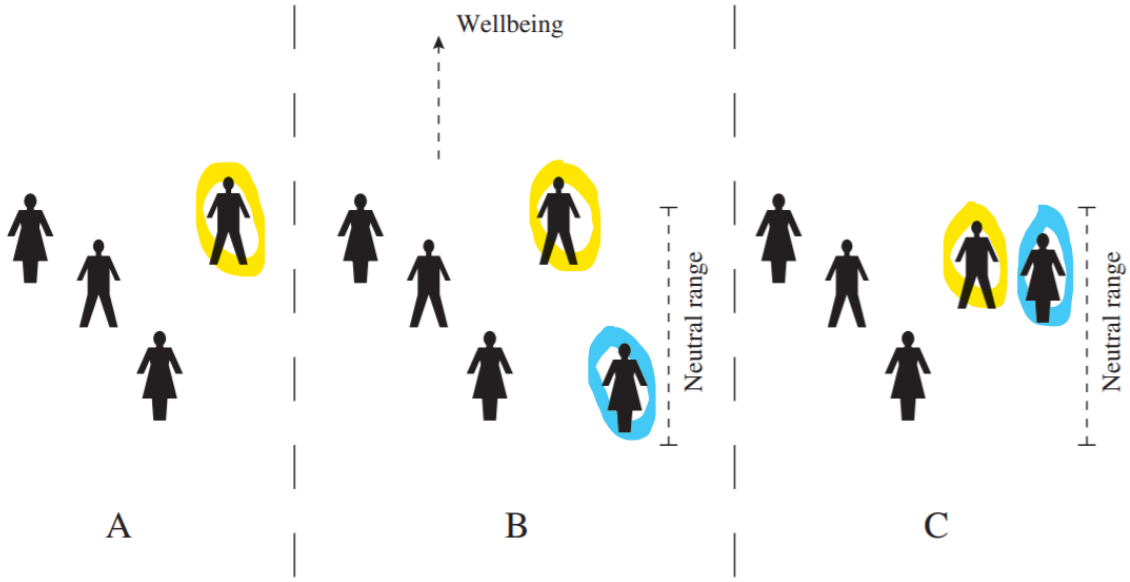

Figure 1: Three worlds depicting existing (yellow) and additional (blue) individuals with differing levels of welfare

Broome claims that applying the IoN.i gives rise to statement (1): 'C is incommensurate with A: C is neither better nor worse than A. It is also not equally good.'

Recall that this statement only holds if:

a) C contains an extra person that A does not

[32] Broome, 2004, p. 401.

b) C is the same as A in all other respects (beyond the addition of the extra person)

C contains an extra person, *blue,* that A does not, meaning we have fulfilled condition (a). However, as *yellow* has a lower level of welfare in C than he does in A, C is *not* the same as A in all other respects. As we have failed to fulfil condition (b), the IoN.i *does not* commit us to saying that A is incommensurate with C.

Ultimately, Broome's assertion that 'adding a person, though neutral, can cancel out bad things too'[33] overstates the IoN.i. Whilst it is true that the IoN.i always claims that *adding* a person to the world is ethically neutral — meaning world A is neither better than, nor worse than C *due to blue's coming into existence* — this notion of neutrality is not overriding in the way that Broome suggests. This is because the IoN.i only commits us to saying that the 'addition [of a life] has no positive or negative value in itself.'[34] As Wodek Rabinowicz articulates, this neutrality need not 'count against other values.'[35] Thus, we can still conclude that C is worse than A — because of the decreased welfare of *Yellow* — whilst also maintaining that the addition of *Blue* is neither good nor bad. As premise (1) relies on an overinterpretation of the IoN.i., we can reject Broome's argument from greedy neutrality as it is unsound.

Overall, the IoN.i only commits us to saying that two worlds are incommensurate provided we fulfil condition **(a) —** one world contains an extra person that the other world does not **—** *and* condition **(b) —** the worlds are the same in other respects. As there is no way to reach Broome's conclusion of 'greedy neutrality' without changing the welfare of some existing people across worlds — i.e without failing to fulfil **(b) —** it seems we have provided a compelling counter to Broome's assertion that the IoN.i can 'neutralise' the bad.

IV.E) Against Incommensurability

The IoN.i commits us to saying that two worlds are incommensurate when **(a)**: one word contains an extra person that the other world does not; and **(b):** the worlds are the same in all other respects. However, Broome questions this incommensurability interpretation, contending that we *must* adopt an equal goodness interpretation of the IoN. This follows from Broome's assumption that the betterness relation is complete, meaning two distributions being compared *cannot be* incommensurate. As Broome explains:

> My assumption is that, if we pick any two distributions in [the field of the betterness relation] then either one of them is better than the other, or the other is better than the one, or the two are equally good. This assumption rules out incommensurateness within the field of betterness. I shall call it the assumption of completeness.[36]

However, this assumption of completeness is not well-founded. Recall the artist example in ***IV.a,*** which suggests it is impossible to ascertain whether Mozart or Picasso is the better artist, as the two artists are so distinct that their work lacks a common measure of comparison. This example illustrates that Broome's assumption of completeness is incorrect, as two things within the field of betterness *can* be incommensurate if the things being compared

[33] Broome, 2005, p. 409.
[34] Broome, 2004, p. 146.
[35] Rabinowicz, 2009, p. 203.
[36] Broome, 2004, p. 22.

are suitably qualitatively distinct. Thus, if we accept that the two distributions being compared in example **Y — 1A** and **1B/1C** —*are* suitably distinct, we can maintain that **1A** is incommensurate with **1B/1C**, thereby overcoming Broome's objection from transitivity.

Admittedly, it is not self-evident that **1A** and **1B/1C** *are* suitably distinct. As Frick articulates, 'prima facie, it is not clear that a mere difference in size between populations makes for an important qualitative difference.'[37] Frick attempts to explain *why* it is not arbitrary to hold **1A** and **1B/1C** to be incommensurate by referring to the conditional value of wellbeing (CVW). Frick advances that when we consider the value of human wellbeing, most of us feel that increasing welfare is valuable *because* humans are *themselves valuable*. Few would attest that people are valuable as a means of promoting human welfare, as this implies that wellbeing is valuable *in and of itself.* As Frick affirms 'human wellbeing matters because people matter – not vice versa.'[38] Thus, Frick advances that 'human wellbeing has contributory value, but this value is conditional on the existence of those to whom it accrues.'[39] Whilst the wellbeing of *existing* people is important —meaning we have an obligation to increase the welfare of existing people when possible — there is no inherent value in human wellbeing that obtains independent of existence, meaning we have no obligation to produce lives *in order to* increase wellbeing. In other words, there is 'no unconditional contributory value in creating further happy lives ...[meaning] adding new happy people to the world is axiologically neutral.'[40]

Whilst the CVW offers a highly plausible account of the way we value human wellbeing and future lives, it is unclear how it provides — as Frick claims it will — a 'deeper motivation for the thought that the Intuition of Neutrality might be best cashed out in terms of incommensurateness.'[41] Frick's argument, though sound, ultimately fails to address our fundamental concern regarding the IoN.i. Namely, the rationale underlying *why* it is not arbitrary to hold **1A** and **1B/1C** to be incommensurate.

Indeed, it is not clear why **1A** and **1B/1C** are incommensurate according to *any* of the orthodox rationales outlined in ***IV.a.*** Firstly, they do not contain incommensurate values, as in both instances we are comparing the same value: human welfare. Likewise, they do not lack a common point of comparison that combines all relevant considerations, as we are simply comparing human welfare across worlds. Finally, they are not suitably distinct, to the extent that a comparison between their goodness is impossible. Whilst they are *quantitively* distinct — as **1B** and **1C** contain an individual, Grace, that **1A** does not — it remains to be shown how this could constitute an important qualitative difference. Indeed, it is unclear how we *could* develop any such rationale without contradicting the very neutralist intuition that motivated us to adopt the IoN.i in the first place. This is because any explanation that suggests **1A** and **1B/1C** are different in some important qualitative sense would thereby imply that there is something qualitatively salient — i.e *valuable* — about the mere addition of Grace. However, the IoN.i is committed to the exact *opposite* intuition: that adding an extra person to the world is axiologically neutral.

Evidently, the IoN.i has reached somewhat of an impasse.

---

[37] Frick, 2017, pp. 16-17.
[38] Ibid, p. 17.
[39] Ibid, p. 18.
[40] Ibid, p. 19.
[41] Ibid, p. 17.

IV.F) THE IoN.i: SOUND OR SENSELESS?

This section has defended an incommensurate interpretation of the IoN that states that 'Ceteris paribus, the world with added people at wellbeing levels within the neutral range is incommensurate with the world not containing these people.'[42] Whilst the IoN.i is initially appealing because it allows us to overcome Broome's objection from transitivity; it is ultimately unable to explain *why* we hold certain populations to be incommensurate.

Consequently, we are left with two options:

Firstly, we can bite the bullet and accept the IoN.i on purely arbitrary grounds. Thus, we would concede that whilst there is no orthodox rationale *why* **1A** and **1B/1C** are incommensurate; we can nevertheless accept that they *are* incommensurate, meaning **1A** is neither better than, nor worse than, nor equally as good as **1B/1C**. Secondly, we could reject the IoN.i, arguing that — until we can explain *why* **1A** and **1B/1C** are incommensurate — the IoN.i is ultimately not tenable.

If you are convinced by the first option, we can accept the IoN.i and its fundamental supposition that adding people to the world is ethically *neutral*. Thus, we can reject Parfit's AAL as we have shown that the assumption on which it rests — that there is disvalue in failing to add future lives to the world — is false.

However, if one is more swayed by the second option, we cannot refute the AAL, as the plausibility of the IoN.i is yet to be confirmed. Indeed, in order to use the IoN.i to refute the AAL, we would first have to provide a coherent rationale as to *why* **1A** and **1B/1C** are incommensurate. However, as I have already articulated, I am not sure that it is even *possible* to provide such a rationale. Therefore, unless we are content to accept the IoN.i on purely arbitrary grounds, I think it is fruitless to try and reject the AAL by using the IoN.i.

Therefore, the following section presents an alternative defence against the AAL. If you believe that we can arbitrarily accept the IoN.i and so reject the AAL, you can proceed straight to the conclusion. If you are not swayed by the IoN.i, you may find the following rationale more compelling.

V: AN ALTERNATIVE RATIONALE AGAINST THE AAL

The AAL relies on the assumption that there is value in adding future potential lives to the world and, consequently, disvalue in failing to add those lives. This assumption does not seem well-founded. **Sections III** and **IV** attempted to explain the shortcomings of this assumption through the IoN. However, even the most plausible interpretation of the IoN — the IoN.i — rests on somewhat arbitrary foundations, and so does not provide a robust defence against the AAL.

This section attempts to provide an alternative explanation as to *why* adding lives to the world is ethically neutral. Ultimately, if this rationale is compelling, we can refute Parfit's AAL by showing that its central assumption — that failing to add future lives to the world generates a large amount of disvalue — is misguided.

[42] Rabinowicz, 2009, p. 4.

V.A) The Conditional Value of Wellbeing

Recall Frick's account of the CVW, which holds that 'human wellbeing has contributory value, but this value is conditional on the existence of those to whom it accrues.'[43] As there is no inherent value in human wellbeing that obtains independent of existence, there is 'no unconditional contributory value in creating further happy lives ...[meaning] adding new happy people to the world is axiologically neutral.'[44]

Although Frick initially presented the CVW to defend the incommensurate interpretation of the IoN, it provides a compelling counter to the AAL in its own right. If the CVW is correct, and human wellbeing is *conditional* on the existence of those to whom it accrues, there is no *unconditional* value in creating future lives. Consequently, there is no unconditional *disvalue* in failing to create those lives. Thus, the CVW — if tenable — allows us to refute the AAL by undermining its central assumption: that failing to create future lives generates a large amount of disvalue.

Initially, the CVW *does* seem highly credible, as it helps explain many of our intuitions outlined in ***III.a*** regarding the value of future lives. Firstly, it explains our intuition that, even if the existence of a child is ethically neutral, the coming into existence of that child can create a better state of affairs if they create positive *instrumental* value for existing people. The CVW reveals that this state of affairs is better because the value of wellbeing is conditional on *existence,* meaning 'adding wellbeing to the world by increasing the happiness of existing people makes the world go better.'[45] Likewise, it coheres with our intuition regarding procreation asymmetry: whilst we have no obligation to create happy people, we do have an obligation *not* to create very unhappy people, as 'adding a person whose life is not worth living makes the world go worse.'[46]

One could retort that whilst the idea of procreation asymmetry is intuitive, it is ultimately incoherent. Surely, if there is *disvalue* in adding a *negative* life, the reserve is also true, meaning there *is* value in adding positive lives to the world. Nevertheless, an analogy with keeping secrets shows that the notion of value asymmetry is highly plausible.[47] Imagine my best friend, Ava, tells me a secret. This secret is incredibly personal, and Ava asks me not to tell it to anyone. Clearly, it is *valuable* if I follow Ava's wishes and take the secret with me to my grave. Moreover, as the value of keeping this secret is conditional, this value only obtains once the secret has been told. In other words, Ava's decision to tell me the secret is not *unconditionally valuable,* meaning her decision to tell me the secret is *ethically neutral.*

Conversely, if I *know* that when I hear the secret, I am likely to tell everyone the contents of her secret, it appears I have a strong reason_m to convince Ava *not* to tell me the contents of her secret, as I know the outcome will be bad if she does. However, I have no reason_m to try and get Ava to tell me a secret if I know I'll be able to keep it; as already elucidated, being told a secret is not in itself valuable, even if we value the keeping of a secret. As this analogy shows, we can have a strong reason_m not to do something if we know the outcome will be bad, whilst having no corresponding reason_m to try and promote something if

[43] Frick, 2017, p. 8.
[44] Ibid. p. 19.
[45] Ibid.
[46] Ibid.
[47] Inspired by Frick's analogy with promise keeping.

the outcome could be good. This is the same rationale underpinning the concept of procreation asymmetry.

Furthermore, the promise-keeping analogy elucidates *why* there is no unconditional value in creating a life *even if* this life would be very worth living. Assuming that 'keeping a secret' is analogous to 'a life that is worth living', and 'deciding to tell a secret' is akin to 'creating a life that is worth living'; this analogy demonstrates that we have no obligation to create a life *even if* this life is worth living. As the CVW asserts, human life is only valuable *once it comes to exist;* the mere creation — or coming into existence — of a life is ethically neutral, meaning we have no obligation to create lives. Indeed, this allows us to advance an even stronger account of neutrality than the one outlined by the IoN. Recall that the IoN states that 'adding a person P to the world is ethically neutral' *provided* that person P would have wellbeing within the 'neutral range', in which 'nonexistence is equally as good as [P's] living with this level of wellbeing.'[48] Consequently, if future people, P, have levels of welfare that are above this neutral range, the IoN no longer concludes that adding P to the world is ethically neutral. However, as Bostrom theorises, considering developments in technology and healthcare, it is not untoward to suggest that future people *will* have exceptionally high standards of living, to the extent that the quality of their lives could be far above the neutral range.[49] Thus, to refute the AAL, we need an account of neutrality that allows us to conclude that we are not obligated to create future people, P, *even if* P would have very high levels of welfare; the CVW allows us to do exactly that. The CVW states that the value of P's wellbeing only obtains *once* P comes to exist. Thus, even if P has such a high level of welfare that P's life would be *extremely* worth living, the value of this wellbeing is still *conditional* on P's existence. Consequently, the CVW states that the mere *addition* of P remains neutral, *even if* P's wellbeing level would be significantly better than non-existence. As John Broome and Adam Morton quip 'people are valuable but creating them is not.'[50]

Overall, the CVW offers a highly plausible account of the way we value populations; and allows us to conclude that creating a person, P, is axiologically neutral *even if* P would have a level of welfare above the neutral range. This means that there is no disvalue in failing to *create* a future person, P, as the value of P's welfare is c*onditional* on P's coming into existence, and so cannot count in favour of *bringing P into existence.* Consequently, the CVW, if correct, undermines the AAL by showing that its underlying assumption is flawed.

V.B) Revisiting Example Y

To test the credibility of the CVW, I will consider what implications the CVW has regarding example Y: the same example Broome proposed to 'test' the validity of the IoN.

Recall example **Y:**

$$1A: (w_1, w_2, \ldots, w_n, \Omega)$$
$$1B: (w_1, w_2, \ldots, w_n, 1)$$
$$1B: (w_1, w_2, \ldots, w_n, 2)$$

According to the CVW, which world is better?

[48] Broome, 2004, p. 140.
[49] Bostrom, 2013.
[50] Broome and Morton, 1994, p. 197.

The CVW states that the value of human wellbeing is *conditional* on the existence of those to whom it accrues. As Grace exists in 1B and 1C, the CVW suggests her wellbeing in those worlds has value, meaning her wellbeing should be considered in any comparison of the goodness of each population. As Grace has higher welfare in 1C than she does in 1B, and the worlds are identical in all other respects, it follows that 1C >1B. Likewise, the CVW states that the coming into existence of a person is *ethically neutral*. As Grace's coming into existence is neutral, 1B and 1C can be no more valuable than 1A due to the *mere addition* of Grace. Consequently, the CVW suggests that 1A is equal to both 1B and 1C.

This conclusion violates the principle of transitivity, meaning that the CVW falls privy to the same 'objection from transitivity' levied against the IoN in ***III.b****:* if 1A=1B, and 1A=1C, the principle of transitivity states that 1B must be equal to 1C. As the CVW suggests that 1C > 1B, it violates the principle of transitivity. Thus, if Broome is correct that 'the relation 'equally as good as' is transitive by a matter of logic',[51] the CVW must be false for implying that it is not.

The following subsection considers two accounts of goodness to ascertain whether Broome is correct that the 'equally as good as' relation is necessarily transitive.

V.C) Two Accounts of Goodness

Broome asserts that the betterness relation is transitive '*by a matter of logic'.* Whilst this is true according to one account of goodness, the Internal Aspects view, the following subsection reveals that if we adopt a *Comparative* view of goodness, the betterness relation is *not* necessarily transitive.

The Internal Aspects Account of goodness (IAG) states that 'roughly, for each outcome, O, how good that outcome is all things considered depends solely on how good it is with respect to each moral ideal that is relevant for assessing the goodness of outcomes, and on how much all of the relevant ideals matter vis-à-vis each other, where these depend solely on O's internal features'.[52] In other words, the goodness of an outcome, O, is determined *solely* by features that are internal to O itself, meaning external features and alternative outcomes have no bearing on how good we value O to be. If the IAG is correct, it follows that 'the relevance and significance of the factors for determining an outcome's value will not vary depending on the alternative with which the outcome is compared [meaning that] the 'all-things-considered better than" relation will be transitive.'[53]

Conversely, the Comparative Account of goodness (CAG) states that 'there is at least one outcome, O, such that there is no answer to the question of how good O is all things considered based solely on O's internal features.'[54] Thus, the CAG rejects the orthodoxy that the goodness of an outcome is determined *solely* by its internal features, suggesting instead that the goodness of O is an essentially comparative relation that is determined by the other outcomes that O is being compared to. Consequently, the goodness of an outcome, and the way it ranks against other outcomes, can change depending on the context of comparison. For example, a pairwise comparison of the goodness of two outcomes — A and B — may not hold once we compare A and B with another outcome, C. Similarly, whilst world A may be equal

[51] Broome, 1994, p. 170.
[52] Temkin, 2012, p. 371.
[53] Ibid, p. 230.
[54] Ibid, p. 372.

to both B and C during a pairwise comparison — of A and B, or A and C — it does not necessarily follow that B is equal to C.

If the CAG is correct, it has two revisionary implications for the way we compare outcomes. Firstly, it suggests that the principle of the Independence of Irrelevant Alternatives (IIA) is false. The IIA states that 'however we should rank A and B in comparison with each other if they were our only two alternatives, that is how we should rank A and B in comparison with each other if they were among a set with any number of alternatives. In other words, neither adding to nor subtracting from a set of alternatives whose members include A and B will affect how A compares to B.'[55] As the CAG states that the goodness of two outcomes, A and B, is *determined* by the alternatives to which they are being compared, the CAG contradicts the IIA as it suggests that alternative outcomes are *highly* relevant for the way we compare outcomes. Secondly, if the CAG is correct, it means that the "all things considered better than" relation may be nontransitive.'[56] Indeed, if the goodness of an outcome is determined by the outcomes to which it is being compared, it follows that the betterness relation between different outcomes will not hold once the context of comparison changes. In other words, the betterness relation will be *nontransitive* across different contexts of comparison.

If the CAG is correct, we can reject Broome's assertion that the 'equally as good as' relation is nontransitive. Consequently, we can accept the CVW. Indeed, on closer inspection, it's clear the CVW *relies* on the CAG. The CVW advances that worlds 1B and 1C could not be more valuable than 1A due to the *mere addition* of Grace, because the addition of future lives is ethically neutral. The caveat 'due to the mere *addition* of Grace' reveals that the CVW's evaluation of 1A and 1B/1C's goodness is essentially *comparative*, as it is determining the value of Grace's welfare in worlds 1B and 1C *by comparing it to* world 1A, in which Grace does not exist. However, this 'equally as good as' relation does *not* hold when we compare 1B with 1C. As the CAG explicates, the goodness relation changes along with the context of comparison, meaning the results of a pairwise comparison may not hold when we compare these worlds with another outcome. If the CAG is correct, and the goodness of an outcome is *determined* by the outcomes to which it is being compared, we can advance that 1B/1C are equal *when compared to 1A*, whilst also maintaining that 1C > 1B *when compared against each other*.

Evidently, then, the CVW *relies* on the veracity of the CAG. If we can prove the plausibility of the CAG, it follows that we can accept the validity of the CVW; the following subsection considers the CAG in more depth to ascertain whether it is a plausible account of goodness.

V.D) Examining the Plausability of the CAG

Larry Temkin (2012) suggests that there are two accounts of goodness: the CAG and the IAG. The IAG suggests that the goodness of an outcome is determined by the outcome's internal features. Consequently, it suggests that the goodness relation is transitive. As the principle of transitivity is widely considered to be a fundamental principle of practical rationality[57], the IAG is the most appealing and widely accepted account of goodness because of its ability to preserve the transitivity of the better-than relation. Conversely, the CAG may

[55] Ibid. p. 388.
[56] Ibid. p. 23.
[57] Ibid. p. 364.

appear less plausible because it implies that the betterness relation is nontransitive. In this subsection, I advance that this implication, whilst revisionary, is not as implausible as it initially appears, meaning it should not count against the plausibility of the CAG.

The CAG implies that, when comparing different outcomes, the betterness relation is nontransitive. Although some may view this as a weakness of the account, an analogy with permissibility reveals that nontransitivity is actually a highly *plausible* and intuitive implication of the way we compare options.

Imagine we have promised our friend, Amelia, that we will cook her dinner. We thus have a duty, D,[58] to fulfil our promise and cook Amelia dinner. However, on the way to Amelia's house, we see a child being attacked by a rottweiler; saving the child from the rottweiler would constitute an act of supererogation, S. Few would dispute that when confronted with a choice between performing D and S, we are permitted to choose S over D, thereby saving the child but, in doing so, failing to cook Amelia dinner. Equally, when confronted with the choice of performing S, and performing a self-interested act SI — such as going bowling with friends — it seems we are permitted to perform SI over S. Whilst it would undoubtedly be *good* for us to save the child, we are not *required* to do so. Indeed, as saving the child would put us at a high risk of being attacked ourselves, S is a supererogatory, not *obligatory,* act. Therefore, it is permissible for us to choose SI over S. However, when confronted with the option to commit all acts – D, S and SI – it would be *impermissible* to choose SI. Whilst it would still be permissible to choose S over D, it is not permissible to choose a self-interested act *over* a duty. As Temkin summarises, this illustrates that 'permissible to do rather than' is a nontransitive relation, since there might be three actions SI, S, and D, such that it might be permissible to do SI rather than S, and S rather than D, and yet not be permissible to do SI rather than D.'[59]

Evidently, nontransitivity is a highly *intuitive* implication when it comes to comparing the permissibility of performing different actions. As Temkin affirms, 'where different factors are relevant for comparing alternatives, transitivity can fail to hold, as it does in this case.'[60] Furthermore, this analogy illustrates that the IIA is *not* a plausible concept when it comes to the permissibility of choosing different outcomes. As Temkin affirms:

> It has long been recognized that whether or not we ought, morally, to do something will depend on its alternatives. …Some alternatives are morally more compelling than others, [meaning] one ought not to do a less compelling alternative, when a morally more compelling one is available.[61]

It is thus a strength, not a weakness, of the CAG that it contradicts the IIA.

Assuming that the betterness relation and the permissibility relation are suitably alike, this analogy is also enough to demonstrate that the betterness relation can be nontransitive. As Temkin affirms, 'the obligatoriness relation may be nontransitive for the same underlying reason that the "not worse than" and "permissibility" relations are.'[62] Thus, if permissibility is nontransitive because the obligatoriness of an action is dependent on the alternatives to which it is being compared, it follows that the betterness relation can be nontransitive if the goodness

[58] Example based on Temkin (2012, p. 195) and Kamm (1995, p. 118).
[59] Temkin, 2012, p. 195.
[60] Ibid. p. 196.
[61] Ibid. p. 202.
[62] Ibid, p. 201.

of an outcome is dependent on the alternatives to which it is being compared; this is exactly what the CAG asserts.

Overall, it seems *highly plausible* to suggest that goodness is *comparative,* meaning the betterness relation is nontransitive. Broome is thus incorrect that we should reject the CAG because it violates the principle of transitivity. Consequently, the CAG appears to be a tenable account of goodness; as the CVW relies on the veracity of the CAG, this allows us to accept the CVW. Therefore, we can conclude, in line with the CVW's central assertion, that creating a future life, P, is axiologically neutral as the value of P's welfare is conditional on P's existence. This undermines the AAL, as the CVW demonstrates that its core assumption — that there is disvalue in failing to add future lives to the world — is misguided.

V.E: The Final Hurdle

The previous subsection concluded that we can accept the CAG and, consequently, the CVW it gives rise to. In doing so, we can conclude that adding future lives to the world is not valuable, thereby refuting the AAL.

However, one can reject this conclusion by arguing that it remains to be proven that the CAG is a fully tenable account of goodness, meaning we cannot accept the CVW and its assertion that creating future lives is not valuable. Indeed, we merely assumed that if the betterness and permissibility relations are suitably akin in nature, the nontransitivtiy of the permissibility relation is *sufficient* to prove the nontransitivity of betterness relation and, consequently, the veracity of the CAG. Although I believe that this *is* a plausible assumption, I concede that because that permissibility relation is deontic, and the betterness relation is axiological, it is not clear that they *are* suitably alike. Consequently, it does not *necessarily* follow that the betterness relation is nontransitive because the permissibility relation is nontransitive. If one is swayed by this objection, we are unable to accept the CAG and the CWW it gives rise to.

However, *even if* we are forced to abandon the CVW, thereby conceding that creating future lives is valuable, we can still claim that we are not *obligated* to mitigate existential risks. Recall that we are not obligated to perform a supererogatory act over a self-interested act/duty; if mitigating existential risks can plausibly be deemed a supererogatory act, it follows that we are not *obligated* to mitigate existential risks *even if* it is valuable to create future lives. This seems coherent in the same way that we are not *obligated* to choose S over D, *even though* it would be in some sense *better* if we did choose S, as we would save the child from being attacked by the dog. As most proposed mitigation policies involve diverting huge sums of money to space exploration, or fundamentally changing our way of life so as to drastically reduce the effects of climate change,[63] I believe that it *is* plausible to suggest that mitigating existential risks *is* a supererogatory act, as it involves drastic measures that surely go beyond mere 'duty'.

Unless there is some way to mitigate existential risks that does not involve going far beyond our duty, this argument from supererogation is enough to prove that we have no obligation to mitigate existential risks.

[63] Bostrom, 2013.

CONCLUSION

In this article, I have attempted to refute Parfit's AAL to show that the consequences of an existential event are nowhere near as catastrophic as the AAL assumes. If the AAL is incorrect, this greatly reduces our reason to mitigate existential risks from the **strong reason$_m$** to the **intuitive reason$_m$**.

I have proposed two ways to reject the AAL: the IoN.i and the CVW.

The IoN.i states that '*Ceteris paribus, the world with added people at wellbeing levels within the neutral range is incommensurate with the world not containing these people.'* As the world is not *better* due to the addition of people with lives within the neutral range, it follows that creating future people is not intrinsically valuable, meaning there is no *disvalue* in failing to create future lives.

The CVW states that human wellbeing is conditional on the existence of those to whom it accrues. Consequently, it suggests there is no unconditional value in creating future lives and, correspondingly, no disvalue in failing to create lives *even if* those lives would have wellbeing above the neutral range.

Both the IoN.i and CVW undermines the AAL by showing that its central assumption – that there is disvalue in failing to create life- is incorrect. Thus, the AAL is incorrect that an existential event would be *catastrophic* because of the disvalue created by all the future people who will never come to exist if humanity goes *prematurely* extinct. Consequently, we can reject the **strong reason$_m$**, thereby greatly reducing our reason$_m$ to mitigate existential risks to only the **intuitive reason$_m$** . Applying principle **(Pd*)** — the strength of our reason to mitigate existential risks is a function of the severity of the consequences of an existential event — gives rise to my **intuitive conclusion**:

> *Whilst we have an **intuitive reason$_m$** to mitigate existential risks to humanity[64], the extinction of humanity is not a severe enough consequence to generate a very strong reason, or **obligation,** to mitigate existential risks.*

One could retort that the consequences of humanity going extinct *are* catastrophic enough to generate an *obligation* to mitigate existential risks. However, if an existential event is only bad because it causes the *existing population* to die; it does not matter from an axiological perspective whether humanity goes extinct today or in a million years' time, as *any* existential event will cause the existing population to die. As the CVW affirms, as there is no unconditional value in creating future lives, we are not required to make humanity endure for as long as possible, meaning there is nothing bad about the *premature* extinction of humanity. And as James Lenman advances 'it is inevitable that our own species will only endure for a finite time…meaning [extinction] is a fate awaiting some generation or another.'[65] Therefore, if the extinction of humanity is *inevitable,* there is *nothing* we can do to prevent humanity from going extinct; regardless of what mitigation policies we pursue, we cannot prevent humanity from going extinct *at some point.* Surely, then, we cannot have *any* reason$_m$ to mitigate existential risks to prevent humanity from going extinct, as this implies that we *ought* to try and prevent the inevitable. As Kant affirms, 'the action to which the "ought" applies must

[64] In order to stop humanity going extinct.
[65] Lenman, 2002, p. 254.

indeed be possible under natural conditions.'[66] As it is impossible to stop humanity from going extinct *eventually*, we cannot have any $reason_m$ to try and prevent it.

This gives rise to my **strong conclusion:**

> *We do not have any $reason_m$ to mitigate existential risks to humanity; because it is inevitable that humanity will go extinct eventually, meaning it is impossible to prevent humanity's extinction, and we cannot have a $reason_m$ to attempt the impossible.*

Although this **strong conclusion** may seem unpalatable, I believe it is the most plausible conclusion regarding our obligation, or lack thereof, to mitigate existential risks. However, the **strong conclusion** relies on CVW and the CAG on which it rests. Therefore, if you did not find my defence of the CAG in ***V.d*** compelling, you could reject the **strong conclusion** by arguing that the principles underpinning it are unsound. However, even if you did not find my defence of the CAG compelling, and so do not believe we can accept the CVW and refute Parfit's AAL, you should still accept my argument from supererogation outlined in ***V.e***: as mitigating existential risks would involve diverting large sums of money to mitigation policies, or radically changing our current way of life, it is *supererogatory*. This gives rise to my **weak conclusion**:

> *Even if creating future lives is valuable, meaning we have some $reason_m$ to mitigate existential risks to humanity, mitigating existential risks is a supererogatory act, meaning we are nevertheless not obligated to mitigate existential risks.*

Therefore, even if you reject my **strong conclusion** — maintaining that we still have *some* $reason_m$ to mitigate existential risks — you should still accept, in line with my **weak conclusion**, that mitigating existential risks is a *supererogatory* act. Thus, even the weakest of my three conclusions affirms that we are not *obligated* to mitigate existential risks.

REFERENCES

Bostrom, Nick. (2013). 'Existential Risk Prevention as Global Priority'. *Global Policy*, Vol. 4, No. 1, pp.15-31

Bostrom, Nick. (2001). 'Existential Risks - Analyzing Human Extinction Scenarios and Related Hazards'. *Journal of Evolution and Technology*, Vol. 9, No.1, pp. 1- 36

Broome, John. (2005). 'Should We Value Population?'. *The Journal of Political Philosophy*, Vol. 13, No. 4, pp. 399-413.

Broome, John. (2004). *Weighing lives.* Oxford: Oxford University Press

[66] Kant, 1998, p. 473.

Broome, John. and Morton, Adam. (1994). 'The value of a person'. *Proceedings of the Aristotelian Society, Supplementary Volumes,* Vol. 68, pp. 167 – 185

Chang, Ruth. (1997). "introduction'. *Incommensurability, Incomparability, and Practical Reason,* R. Chang (ed.). Cambridge: Harvard University Press

Frick, Johan. (2017). On the survival of humanity. *Canadian Journal of Philosophy,* Vol. 47, No. 2, pp. 344-367

Frick, Johan. (forthcoming). 'Context-Dependent Betterness and the Mere Addition Paradox'. *Ethics and Existence: The Legacy of Derek Parfit,* McMahan, J., Campbell, T., Goodrich, J. and Ramakrishnan, K. Oxford: Oxford University Press

Gert, Joshua. (2012). 'Moral Worth, Supererogation, and the Justifying/Requiring Distinction'. *The Philosophical Review,* Vol. *121,* No. 4, pp. 611-618

Hsieh, Nien-He. (2016). "Incommensurable Values". *The Stanford Encyclopedia of Philosophy,* Zalta, E. (ed.)

Kamm, Frances. (1985*) Supererogation and Obligation*. The Journal of Philosophy, Vol. 82, No. 3. pp. 118-138

Kant, Immanuel. (1998). *A Critique of Pure Reason*. Cambridge: Cambridge University Press

Lenman, James. (2002). 'On becoming extinct'. *Pacific Philosophical Quarterly,* Vol. 83, pp. 253 – 269

Narveson, Jane. (1973) 'Moral Problems of Population'. *The Monist,* Vol. *57, No.* 1, pp.62-86

Ord, Toby. (2020). *The Precipice: 'a Book That Seems Made for the Present Moment'.* New Yorker: Bloomsbury Publishing

Parfit, Derek. (2016). 'Can We Avoid the Repugnant Conclusion?'. *Theoria,* Vol. 82, No. 2, pp.110-27.

Parfit, Derek. (1984). *Reasons and persons*. Oxford: Clarendon Press

Rabinowicz, Wlodek. (2009). 'Broome and the Intuition of Neutrality'. *Philosophical Issues,* Vol. 19, No.1, pp.389 – 401

Temkin, Larry. (2012). *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning.* Oxford: Oxford University Press

Temkin, Larry. (1987) "Intransitivity and the Mere Addition Paradox". *Philosophy & Public Affairs, Vol. 1*6, No. 2, pp. 138-187

Wiggins, David (1997). "Incommensurability: Four Proposals." . *Incommensurability, Incomparability, and Practical Reason,* R. Chang (ed.) Cambridge: Harvard University Press

Explaining the Rational Irrationality of Addiction: A Theory of the Self as Causal Explanation

ALICE PESSOA DE BARROS
McGill University

BOTH IN THE MEDIA and in scientific literature, addiction is mostly described as an impulsive, relapsing, psychiatric condition.[1] Studies have shown that at the core of most (dysfunctional) addicts' lives is an endless self-destructive behaviour, as their substance dependence results in a loss of meaningful relationships and a disinterest in other activities.[2] For this reason, addiction is often believed to be a clear manifestation of irrational behaviour, as addicts usually attempt to quit their whole life. This desire to quit possibly indicates a hint of rationality in addicts' desire to change their behaviour; it therefore seems like addicts have the right, rational desires, but are unable to turn them into long-lasting, rational actions. It is this dichotomy between irrational actions and rational desires that makes necessary the attempt to replace a theory of rationality by an alternative model of causal explanation in the case of addiction. A theory of rationality, for instance, would be unable to explain why certain substance users still manage to live fulfilling and balanced lives, as it would systematically link substance use with a failure in a rationalizing mechanism.

In this paper, I will therefore argue that substance use disorders (or SUDs) involve as much rationality as irrationality, thus making it unfruitful to try to make it fit into an explanatory rationality model. I will then discuss possible alternatives to a theory of rationality which could resolve this 'rational irrationality' involved in addiction. Indeed, it may be more useful to consider a theory of self-control or value dysregulation instead, as a way to understand addiction. These two alternative models will be interesting as they will involve looking at addicts' sense of self, how they apprehend their future and what they value in life—I will argue that a comprehensive model of addiction must include this role of selfhood in substance abuse, as it is at the core of their decision-making, and thus at the core of their difficulty to quit. By 'selfhood', we will see that I include any perception that one has of themselves, what they value, and how they predict their behavior based on this perception and value-judgment; for addicts, this often entails being unable to conceive of themselves as non-addicts or finding value in anything other than drugs in their life. This considerably impairs their ability to make decisions that could help them work towards a sense of self that does not revolve around drugs. Finally, we will look at what these alternative theories imply for an assessment of addiction as

[1] Gene M. Heyman "Addiction and Choice: Theory and New Data" *Frontiers in Psychiatry* 4, no. 31 (2013): 1.
[2] Dan Lubman, Murat Yücel, and Christos Pantelis "Addiction, a Condition of Compulsive bBehaviour? Neuroimaging and Neuropsychological Evidence of Inhibitory Dysregulation" *Addiction* 99 (2004): 1491-1502.

a mental disorder, by attempting to make it coincide with the harmful-dysfunction model proposed by Wakefield.

I. The Rational Irrationality of Addiction

Rationality still remains a divisive topic among philosophers of psychiatry. There is, to this day, no general agreement or universal way to define what is rational or irrational. The liberal view wants to call rational any action that has been done following a process of reasoning, or a process of simply finding reasons for an action. Others claim that rationality needs more than a few justifying reasons, but it still remains unclear what is needed in addition to what would *technically* be considered rational reasoning. A theory of rationality is not meant to replace all existing theories of psychiatry so far in terms of mental illness assessment. It is simply a form of causal understanding,[3] or a way in which we can explain why certain people with a condition do certain things and why others don't. In other words, a failure in a 'system' or 'mechanism' of rationality could be a way to understand why patients with mental disorders act in a certain way. However, we cannot look at dysfunctions in rationalization without knowing what a functioning rationalization system does. As Dominic Murphy notes in *Psychiatry in the Scientific Image*, an understanding of such functions will necessarily rely on a normative theory, as well as epistemic intuitions.[4] After all, how can we decide whether something is rational or not without even having a clear definition of what would be considered rational in any given situation? This is also important since it is necessary to have an idea of where to cross the line between common, everyday life irrationality experienced by all of us, and pathological irrationality. A way to reconcile this complexity is to discard the possibility of reaching a comprehensive theory of addiction through the assessment of a presupposed 'rationalization mechanism' altogether. This is what I propose to do in this section, by showing that addiction involves as much rationality as irrationality, and thus, that theories of irrationality do not hold explanatory power in the case of substance use disorders.

Although a theory of rationality might be a useful way to look at many psychiatric conditions, it seems like the case of substance use disorders is more complex because, despite engaging in self-destructive behaviour (which is intuitively considered irrational), most addicts wish that they would quit and stop desiring the substance they are addicted to. This desire to quit, although difficult to turn into actions, is what makes addicts rational in some sense. Harry Frankfurt explains such desires in his paper on freedom of will: he explains that there are two types of desires to take into consideration when trying to assess whether a person's desires are rational or not. First-order desires are the most common desires we experience in everyday life; they are, for instance, our desire for a new car, or for a pastry when we walk by a bakery. Second-order desires, however, are more complex, as they involve not only a desire for something but a desire to desire (or not desire) something.[5] Using the previous example, if my first-order desire is to buy a pastry, my second-order desire would be the desire to desire something healthier instead, like a salad. In other words, I can desire a pastry but, at the same time, wish that I could desire a salad instead, because it seems more reasonable or rational as it promotes my health. Both types of desires are common in everyday life, and not everyone

[3] Dominic Murphy, *Psychiatry in the Scientific Image* (Cambridge MA: MIT Press, 2012), 158.
[4] Ibid, 151.
[5] Harry Frankfurt, "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68, no. 1 (1971): 6.

always chooses the most rational desire of the two. But the presence of both of these desires shows some sort of inconsistency even in our personal, everyday desires, and this is why this model can help us understand substance use disorders and its 'rational irrationality'.

Indeed, the majority of addicts suffer, just like everyone else, of the same type of conflicting desires. Their main desire is to take the drug they are addicted to, but most substance users know they should not take it, and therefore do have a second-order desire to not take the drug—in other words, they *wish* that they did not desire the drug. As Harry Frankfurt explains, this is not uncommon and it certainly does not only apply to people diagnosed with a substance use disorder. We all suffer from these sorts of conflicts daily, thus showing that this wish to desire something is not indicative of a failure in rationalization, or at least not one that needs to be pathologized. This affects the way we conceive of the 'irrationality' in substance abuse, as the way in which they lack rationality seems to be more common to all of us. I therefore propose that, if the presence of sensible 'second-order desires' does not necessarily mean that addicts are rational, it does not mean they are completely irrational either. Instead, it would seem like they simply suffer from a conflict between their (second-order) desires and their actions: on the one hand, their desires are rational in that the addict wishes he did not desire the drug. On the other hand, though, the addict still ends up taking the drug and is unable to stop and listen to his rational desire. If a theory of rationality were to be applied to the case of SUDs, how would one know if they are to assess the rationality of an addict's action or of an addict's desires?

Let us imagine we try to explain addiction with a rationality theory: this would hypothesize that addicts suffer from a dysfunctioning rationalizing mechanism on the basis that they are unable to fight their repeated substance use and to reach their goal of quitting. The dysfunction would have to be either generalized, meaning it would affect actions, thoughts and desires altogether, or more specialized, meaning it would affect only one of these. Since we have seen that addicts seem to have correct desires as second-order desires attest, a rationality theory would only concern actions for instance. Maybe, then, such a theory would propose that the issue for addicts is that they lack rationality in the sense that they have a dichotomy between what they desire and the way they act. But after all, many addicts do overcome their addiction and most users of illicit drugs quit before the age of thirty.[6] Additionally, like the previous part on second-order desires suggests, non-addicts have trouble matching their desires and actions as well. Thus, although it would not be completely impossible, it seems counter-productive to attempt to explain addiction with a theory of dysfunctioning rationality.

In his paper "In What Sense are Addicts Irrational?" Howard Rachlin proposes a similar model by arguing that this inconsistency between actions and underlying motivations is precisely what makes addicts irrational. According to him, addiction cannot possibly be rational because, even if the rationalizing mechanism functions properly for addicts, it is their underlying *current* motivations, or 'visceral factors' that make them irrational in the end.[7] He thus argues that one of the reasons why addicts seem unable to resolve this gap between action and motivation or desires is because they fail to predict how they will act in the future. So for instance, if a smoker takes a cigarette while swearing that it is his last one, it shows that he cannot anticipate the fact that it will most likely not be his final cigarette. Because of this constant failure in pattern anticipation, addicts cannot possibly change their behaviour in the

[6] Heyman "Addiction and Choice", 4.
[7] Howard Rachlin, "In What Sense Are Addicts Irrational?" *Drug and Alcohol Dependence* 7, no. 1 (2007): 11

long run, as they are unable to predict it. This inability to see this future pattern and act in a way to derive from it is, according to Rachlin, what makes addicts irrational. Additionally, he adds that non-addicts usually have no difficulty identifying with their future selves, which enables them to predict their future behaviour, and act according to it.[8]

Rachlin's argument to prove the irrationality of addiction can be fairly convincing in the sense that it can surely be seen as some sort of irrationality to fail to anticipate future behaviour, specifically when this behaviour has been repeated for years. However, what this argument fails to consider is the fact that there is a significant ontological gap between being an addict and being a non-addict, which is a gap that most of us do not have to consider when making plans in the future. In other words, Rachlin does not take into consideration the fact that substance users cannot plan to stop consuming drugs in the future because they cannot possibly conceive of themselves as non-addicts, or conceive of anything else that might ever hold as much value in their life. In the same way, if a non-addict decided to become an addict in a few months and had to anticipate what it would be like to live with addiction, they would probably not be able to conceive of such a life, because their current sense of self is defined by various things in their life (their family, their profession, their hobbies), and trying to conceive of a self which mostly values drug use is nearly impossible to conceive of, especially if it requires making decisions for this future self. Additionally, it may be irrational and even harmful to try to predict future behaviors in the case of addiction, as Emily Walsh argues in her paper on cognitive transformations. Following Laurie Ann Paul's framework according to which one cannot make life-changing decisions rationally by assessing the future since, epistemologically, this future experience is inaccessible to us in the present,[9] Walsh argues that doing this in the case of mental illness is actually counter-productive.[10] Although according to Paul, this epistemological gap applies to all of us, it does seem even more irrational to try to make future predictions for addicts specifically, and to ask them to anticipate a behaviour in which their values and sense of self will be completely different.

This is what Hanna Pickard discusses in her paper "Addiction and the Self". She explains that an aspect of drug use that has long been overlooked is the value that drugs have for people who consume them, and the fact that they often represent their whole identity.[11] This view can help us understand better why the main issue for substance users is that they cannot identify with their future self, making it unable to conceive of their identity as someone other than a drug user in the long term. So what does this mean for our attempt to understand and explain substance abuse? First, that Rachlin's argument is pertinent but misses the fact that what he considers to be irrationality might actually be an ontological gap that is particularly difficult to resolve in the case of mental illness. And second, that if we cannot seem to explain why addicts keep taking drugs by looking at whether their actions or beliefs are rational or not, it would seem like we may be able to explain this phenomenon by finding a theory that would focus on their sense of self and self-control. Now that I have discussed the difficulty of explaining addiction with a theory of rationalization, I will introduce an alternative model that could possibly be more productive by relying on a theory of the self and its values.

---

[8] Ibid, 8.
[9] L.A. Paul, *Transformative Experience* (Oxford: Oxford University Press, 2014), 120.
[10] Emily Walsh, "Cognitive Transformation, Dementia, and the Moral Weight of Advance Directives," *The American Journal of Bioethics* 20, no. 8 (2020): 59.
[11] Hanna Pickard, "Addiction and the Self," *Noûs*, (2020): 2.

II. Introducing a Theory of the Self to Explain Addiction

Having established that understanding addiction must involve understanding the sense of self of addicts, we can look at a first way to explain addictive behaviour with a theory of self-control, which will be linked with a second theory, centered on value dysregulation. If there does not seem to be anything dysfunctional about addicts' rationalizing mechanism—as we have seen that most of them can be rational about what they ought to desire—we could consider instead that it is their sense of self that makes them remain substance users. Indeed, it seems like what we need to look at when trying to explain addictive behaviours is not the reason why they start, since nobody starts taking drugs knowing that they will be addicted in the long term, but why they cannot seem to quit.

It is important to establish that a theory of the self will only be able to take into consideration addicts who do not live well with their substance abuse, so it will not simply serve to understand addiction but rather to understand *harmful* addiction. Therefore, before we start discussing these alternative models further, we need to ask what makes drugs so terrible? This is an important question because addiction is a very broad term that encompasses many different types and frequencies of substance use; a comprehensive model of this disorder cannot possibly apply to every single person who is addicted to a substance. After all, it seems like many of us know at least one person who enjoys taking drugs regularly but still lives a fulfilling, successful or responsible life. In these cases, who are we to say that, on the basis of their recurring drug use, this person should change their way of living? A theory of rationality would not be able to differentiate between these two types of drug uses: if, according to this model, addiction comes from a dysfunctioning rationalization process, it cannot explain that some people seem to be living a reasonable life as drug addicts. In other words, it seems necessary here to accept that drugs are only harmful if they make one's life worse. In this sense, a theory of the self, or rather of a loss of self, would be more applicable because what seems to be differentiating an addict with a fulfilling life from an addict whose life revolves around their condition is their sense of self: one manages to go beyond their identity as a substance user and to find value in things other than drug use, while the other is unable to live a life that does not revolve around taking drugs, thus meaning they are unable to find a sense of self and value in any other activity. To sum up, a theory of the selfhood of addicts needs to assume that drugs are not intrinsically bad—losing your sense of self to them is what makes them harmful.

The question we are left with now is: how can we find a model able to explain harmful addiction and what would be its implications? Hanna Pickard argues that when talking about drugs, we need to take into consideration the value that drugs hold.[12] This concept of the value of drugs will inform our theory of the self, as controlling ourselves involves some kind of decision making. Therefore, to accept this theory as a comprehensive model of addiction, we need to look at the possibility of addiction as a choice, or at least, as not entirely a compulsion. It seems that the boundary between choice and compulsion is blurred in the case of many addicts—we will see that this is mostly due to the fact that drugs hold significant value, which would justify why someone might *choose* to take them.

The temporary pleasure one gets from drugs is the reason why people try drugs at least once in their lifetime, and it is one of the reasons why they might remain addicted to them. But broadly speaking, the value of drugs goes way beyond this temporary euphory. Drugs have a

[12] Ibid, 4.

social value: they are usually started as part of a social group, either because the whole group decides to try or as a way to fit into that group. This directly relates to a possible theory of the addicted self: the social group that one is part of or associates with makes up a part of one's identity. Since one's community is closely linked to their sense of self, addicts would lose both if they quit, which is one of the reasons why drugs are so valuable. But this value of drugs is not enough to justify that drugs involve some kind of decision making and are not an act of compulsion. Hanna Pickard states in her paper that studies have shown that in a situation in which addicts had the choice between taking their drug of choice or a certain amount of money, they chose the money.[13] Additionally, she notes that many have attempted to prove that addicts do not get any pleasure from the consumption of their drug of choice once addicted, but this is actually not entirely true, as addicts still get pleasure from it even though, with time, this pleasure decreases slightly.[14]

This shows that addiction, although not entirely a choice, cannot be considered as compulsive behaviour only. A clear distinction between the two is difficult to make in the case of substance use disorders, because it is neither all choice nor compulsion: in a sense, addicts compulsively take something they desire and want to take. Arguably, a theory of the self can help reconcile this paradox; it would then seem that the compulsive part of addiction comes from the fact that the addict's sense of self is so deeply rooted in its main activity, or its core (which is taking drugs) that it turns this activity into something that can be associated with compulsion. It is true that it is quite rare that a non-addict becomes so passionate about an activity that it turns into a compulsion. However, a plausible explanation for this comes back to our sense of self again: the reason we do not focus on one pleasurable action compulsively is because our sense of self does not revolve around this particular action. A non-addict with a fulfilling life will have different hobbies and activities, different people they see, a life which, overall, is balanced between different activities. Addicts whose life revolves around their drug of choice do not have so many options, as their sense of self is not defined by other valuable activities in a significant manner. This is further confirmed by the fact that addiction is more common among people of lower economic and social status, who experience stress and/or have experienced childhood abuse for instance.[15] People in these difficult situations are more likely to struggle finding another source of self-fulfillment beside their addiction: without a stable social or professional situation, and because of stress or trauma, drugs provide an easy access to a pleasurable activity, and are more likely to become a main source of enjoyment. This activity thus becomes the most valued part of the addict's life, who often ends up socially and emotionally isolated, leaving them with a sense of self revolving solely around their addiction.

So what does this entail for our understanding of addiction through a theory of self-control? First, it means the addict's sense of self is definitely something to take into consideration when trying to understand such a behaviour. Second, it means that taking drugs over a long period of time is not a choice per say, but it does involve decision-making and cannot be explained by the weight of compulsion only. Keeping this in mind, a theory of self-control, paired with a theory of value dysregulation, could prove to be more useful than a

[13] Pickard, "Addiction," 6. This may also be a reflection of the fact that drug addicts are more likely to have lower social and economic status. It still shows nonetheless that they are capable of responding to incentives and do not simply give in to compulsions.
[14] Ibid, 4.
[15] Walter Sinnott-Armstrong, and Hanna Pickard. "What Is Addiction?" In *Oxford Handbook of Philosophy and Psychiatry* (Oxford: Oxford University Press, 2013): 852.

rationality theory, as it considers the addict's self-assessment, sense of values, as well as the control (or loss of control) of this self. We can now discuss in this final part what this theory might mean for the assessment of addiction as a psychiatric condition.

III. Implications for a Harmful-Dysfunction Analysis of Mental Illness

It is generally accepted that psychiatric disorders involve some kind of malfunction of a mental process. According to a rationality theory, the malfunction that makes addiction a disorder has to do with our rationalization process. However, we have seen that the issue with this model of causal understanding is that it does not take into account the case of substance users who, despite their addiction, seem to be living fairly healthy and responsible lives. If these addicts really did have a malfunctioning rationalization mechanism, that malfunction could not possibly apply only to their drug use, it would have to affect other areas of their lives too. This does not coincide with drug users living an overall reasonable and fulfilling life, and it also cannot explain casual drug users, who might not be 'addicted' in the sense that they do not experience cravings or withdrawal, but still make the decision of using drugs once in a while. We have also seen that a rationality theory fails to see that the reason addicts cannot seem to quit is not necessarily because of their desires, which are often rational, but because of their difficulty to associate with a future non-addicted self, and to find value in anything other than their drug of use.

We are now left wondering what the implications of a self-control theory would be in terms of the way we consider addiction as pathological or not. A popular way to assess mental disorders is Wakefield's two-stage model, or his 'harmful-dysfunction' analysis.[16] According to him, this model can assess whether a behaviour is a mental illness or not in two stages: first, by looking at whether an organ or system is functioning improperly, and second, by looking at how it affects the possibility of living a fulfilling and flourishing life.[17] The benefit of this model is that it allows for the possibility of non-pathological drug use in the case that this use does not affect one's life negatively. As we have discussed earlier, these are crucial cases to consider, because addiction does not affect everyone in a similar way. The idea that all drug users become instantly addicted and end up miserable and impoverished is common, but it certainly is not the norm in terms of drug use. Actually, most users of a drug do not become addicted, and this is true even in the case of cocaine for instance (which is known to be highly addictive), since only one in six users do develop a dependence.[18] It thus seems like the second stage of Wakefield's model would be compatible with our theory of self-control. We can now see if the 'dysfunction analysis' of his two-stage picture concords with what we have established about addiction.

Looking at the first stage of Wakefield's model, which involves deciding that an organ is not fulfilling its function in a proper way, it appears tedious right away to find how this can relate to addiction. If, as we have seen, an addict's behaviour does not come from a malfunctioning rationalization mechanism, we need to find another organ or system working improperly in order to confirm that, according to the harmful-dysfunction analysis, substance abuse is indeed a disorder. The question then is: is losing your sense of self and values—which

[16] J.C. Wakefield, "Disorder as Harmful Dysfunction: A Conceptual Critique of DSM-III-R's Definition of Mental Illness", *Psychological Review,* 99 (2): 232-247.
[17] Murphy, *Psychiatry in the Scientific Image,* 37.
[18] Ibid, 187.

involves losing the ability to control this self—a dysfunction of some sort? Since we do not know of any system or mechanism in the brain designed specifically to develop this, we would have to rely on normative beliefs about what functions right or not. Furthermore, to know if someone or something is dysfunctional, we need to know how a person or thing would work if it were functional. In the case of self-control, this is difficult to assess as even non-addicts often do not control themselves perfectly. But loss of self-control is a part of a loss of one's sense of self to drug use, so the dysfunction to look at here might be this sense of self issue. Indeed, if having no or little self-control is common and thus maybe could not lead us towards a proper dysfunction that could fit Wakefield's model, losing oneself in a certain activity such as consuming drugs is more likely to attest of some sort of dysfunction. And this loss of self to an activity can be seen in other mental illnesses, such as eating disorders, where one arguably loses oneself to the activity of losing weight and controlling every food intake. So, in substance abuse and maybe even in the case of eating disorders, it seems like losing oneself in something in a way that discards any other activity or responsibilities (in this case, drug consumption) could be fitting to the harmful-dysfunction analysis.

Furthermore, this explains why the debate around this disorder still involves discussions of possible addictions to more common things such as sugar, pornography, or work. While many say that one cannot get addicted to these activities as they do not involve any craving or withdrawal (or at least, not in the same way as with drugs), the criteria that might differentiate a habitual, enjoyable behaviour from a harmful dysfunction could be precisely this sense of self and value. For instance, if a person is so addicted to pornography that they cannot find value in any other activity or control themselves when it comes to prioritizing other responsibilities, and if it affects their life in a negative way (which is usually the case if the former is true), then it would seem like one could fit such a behaviour into the harmful dysfunction analysis and call it pathological. So ultimately, even if it is unsure whether a dysfunction in value-judgement and selfhood could be enough to be considered a symptom of a psychiatric condition, it is nonetheless informative to see that if it were to be considered a dysfunction, then the alternative theories proposed in this paper could fit into a proper explanatory model of mental illness.

To conclude this essay and this final part about the implications of a theory of the self for the assessment of substance use disorders, we can ask what would help addicts overcome this loss of self and values? Keeping in mind what has been said throughout this paper, it seems like what addicts need to slowly abandon or diminish their addiction is to develop valuable activities besides taking drugs. That way, they will be able to develop another sense of self that does not include them being solely addicts.

Overall, I have argued in this paper that a theory according to which addicts cannot quit because their rationalization mechanism is dysfunctional would not be pertinent. Instead, I showed that what might explain this behavior is the fact that addicts (specifically the ones whose life is being ruined by drugs) lose value in anything in life except for drugs, which comes to define their entire sense of self. They are thus stuck in their self-identification as addicts, making it impossible for them to conceive of themselves as non-addicts in the future.

REFERENCES

Frankfurt, Harry G. “Freedom of the Will and the Concept of a Person.” *What Is a Person?* (1988): 127–44.

Heyman Gene “Addiction and Choice: Theory and New Data” *Frontiers in Psychiatry* 4, no.31 (2013): 1-4.

Lubman, Dan, Murat Yücel, and Christos Pantelis “Addiction, a Condition of Compulsive Behaviour? Neuroimaging and Neuropsychological Evidence of Inhibitory Dysregulation” *Addiction* 99 (2004): 1491-1502.

Murphy, Dominic. *Psychiatry in the Scientific Image*. Cambridge, MA: MIT Press, 2012.

Paul, L. A. *Transformative Experience*. Oxford: Oxford University Press, 2014.

Pickard, Hanna. “Addiction and the Self,” *Noûs,* (2020): 1-25.

Rachlin, Howard. “In What Sense Are Addicts Irrational?” *Drug and Alcohol Dependence* (2007): 92-99.

Sinnott-Armstrong, Walter, and Hanna Pickard. “What Is Addiction?” In *Oxford Handbook of Philosophy and Psychiatry,* 851-864. Oxford: Oxford University Press, 2013.

Wakefield, J. C. “Disorder as Harmful Dysfunction: A Conceptual Critique of DSM-III-R’s Definition of Mental Illness”, *Psychological Review* 99, no. 2: 232-247.

Walsh, Emily. “Cognitive Transformation, Dementia, and the Moral Weight of Advance Directives.” *The American Journal of Bioethics* 20, no. 8 (2020): 54–64.

A Kantian Account of Aesthetic Judgements in Literature

YIP SZE KAY
National University of Singapore

LITERATURE FALLS under Kant's account of aesthetics in the *Critique of Judgement* (1790). Hume's 1757 essay 'On the Standard of Taste,' to which the *Critique* responds, focuses on aesthetic judgements of literature, while the *Critique* offers judgements of poetry as a case of aesthetic judgement[1] and puts forward Homer and Wieland as examples of artistic genius.[2] However, although Kant's thought encompasses the literary, literature has remained neglected as a case study for Kantian aesthetics. Rather, contemporary Kantian scholarship has focused on general exegesis of the *Critique* or, as Mark Cheetham argues in *Kant, Art, and Art History*, applied his aesthetics to visual art.[3] This oversight may be more historical than philosophical: Cheetham notes that Kant's aesthetics were first adopted by the Cubist movement,[4] which, I suggest, may have generated a longstanding association between Kant and visual art, leading philosophers to overlook the relationship between literature and Kantian aesthetics.

Therefore, this essay offers an account of Kantian aesthetic judgements in literature to demonstrate that applying his account to specific aesthetic mediums resolves issues in scholarship on the *Critique*. Focusing on literature as a test case for Kant's account, I show that literature resolves two problems for Kantian scholarship. Firstly, I establish that the case of literature adjudicates between competing accounts of Kant's suggestion that art engages the understanding and the imagination in harmonious free play. Secondly, I use the case of literature to explain Kant's argument that aesthetic judgements occur without involving any determinate concept.

In this essay, I divide Kant's account into three elements: aesthetic perception (the qualities of the artwork forming the basis of judgement), aesthetic experience (the subjective experience of encountering the artwork), and aesthetic emotion and judgement (the feeling of pleasure leading to a judgement that the artwork is beautiful). In the first section, I offer a Kantian account of aesthetic perception in literature. Next, I focus on three recent accounts of aesthetic experience as free play, rejecting one and combining the other two into a modified account of aesthetic experience in literature. Finally, I draw on literary criticism to demonstrate—contrary to at least one scholar—that, as Kant himself proposes, aesthetic judgements can take place without requiring concepts.

---

[1] Immanuel Kant, *Critique of the Power of Judgement*, trans. Paul Guyer and Eric Matthews (Cambridge: Cambridge University Press, 2000), 7:65.
[2] Kant 1790, 47:308–9.
[3] Mark A. Cheetham, *Kant, Art, and Art History: Moments of Discipline* (Cambridge: Cambridge University Press, 2001).
[4] Cheetham 2001, 3.

Aesthetic Perception in Literature: Formal Play of Shapes and Sensations

For Kant, the basis of aesthetic judgement lies in the form of the artwork. Kant argues that 'taste is not what gratifies in sensation but merely what pleases through its form,' where form 'is either shape or play: in the latter case, either play of shapes (in space, mime, and dance), or mere play of sensations (in time).'[5] Therefore, for Kant, the basis of judgement lies in our encounter with the formal qualities of the artwork, with the specifics of these formal qualities depending on the medium of the artwork.

I define the formal qualities of literature by drawing on Terry Eagleton's *How to Read Literature*, which suggests that 'form' refers to specifically literary qualities such as 'tone, mood, pace, genre, syntax, grammar, texture, rhythm, narrative structure, punctuation, ambiguity.'[6] Given these formal qualities, I argue that aesthetic perception in literature involves play of sensations and, in some cases, shape and play of shapes. From a Kantian perspective, since we read literature line-by-line, our experience of literature is necessarily temporal: thus, as we read, variations in tone, mood, pace, syntax, texture, rhythm, narrative structure, and ambiguity give rise to a play of sensations. For example, in Eagleton's reading of the opening line to E. M. Forster's *A Passage to India*, he argues that 'the tone of the passage—disenchanted, slightly supercilious, a touch overbred—is that of a rather snooty guidebook.'[7] As he continues reading, he notes,

> If the narrator is detached because he has seen too much, as the tone of the passage might suggest, then two contrary feelings—inside knowledge and a rather lofty remoteness—interestingly coexist.[8]

As Eagleton's eye moves across the page, changes in tone, from disenchantment to detachment, cause changes in sensation in the reader, with the changes themselves creating contrast and interest. Thus, our perception of formal qualities such as tone can create a play of sensations serving as the basis for aesthetic judgement in literature. Additionally, some forms of literature, especially poetry, involve a visual element. Thus, our experience of such literary forms involves shape and the play of shapes. For instance, Eagleton argues that in the first line of John Keats's 'To Autumn,'

> The sheer packedness of the line also arrests the eye. … This sensuous richness is meant to evoke the ripeness of autumn, so that the language seems to become part of what it speaks of.[9]

Therefore, our visual perception of formal qualities such as grammar and punctuation may also serve as the basis for aesthetic judgement.

Eagleton's definition of form in literature further supports that the neglect of literature relative to the visual arts in Kantian scholarship is likely due to historical reasons rather than

---

[5] Kant 1790, 5:225.
[6] Terry Eagleton, *How to Read Literature* (London: Yale University Press, 2013), 2.
[7] Eagleton 2013, 8–10.
[8] Eagleton 2013, 11.
[9] Eagleton 2013, 26.

literary difficulties. Formal qualities such as tone, mood, and genre are shared across different aesthetic mediums, whether visual or textual: for example, we might equally discuss the mood evoked by a Brontë novel or a Rothko painting. These shared formal qualities indicate that literature is as equally a candidate for a Kantian account of aesthetic judgements as the visual arts.

Aesthetic Experience in Literature: Purposiveness through Associative Harmony

Kant argues that the formal qualities of the artwork engage our imagination and understanding in a state of harmonious free play, which leads our judgement to confer upon the artwork a sense of purposiveness. He writes:

> If, then, the form of a given object in empirical intuition is so constituted that the apprehension of its manifold in the imagination agrees with the presentation of a concept of the understanding (though which concept be undetermined), then in the mere reflection understanding and imagination mutually agree for the advancement of their business, and the object will be perceived as purposive merely for the power of judgment.[10]

Kantian scholarship focuses on two issues of interpretation in his account. Firstly, scholars have disagreed on the role of the imagination and the understanding in Kant's argument that the representation of an object for cognition involves 'imagination for the composition of the manifold of intuition and understanding for the unity of the concept that unifies the representations.'[11] Secondly, scholars have disagreed on the meaning of non-conceptual play in Kant's argument that the free play of the understanding and the imagination 'occur[s] without presupposing a determinate concept.'[12]

Regardless, it is clear that for Kant, the formal qualities of an artwork generate an interaction leading to mutual agreement between the imagination and the understanding. This state of free play generates a sense of purposiveness, meaning that the artwork is held to have 'deliberately caused' this interaction through its formal qualities. As Kant writes, the representation of the artwork in our consciousness 'contains a determining ground of the activity of the subject with regard to the animation of its cognitive powers, thus an internal causality (which is purposive) with regard to cognition in general.'[13] Kant emphasises that 'purposiveness' is distinct from 'purpose,' because purposive objects may appear willed without necessarily having been purposefully made: 'purposiveness can thus exist without an end, insofar as we do not place the causes of this form in a will, but can still make the explanation of its possibility conceivable to ourselves only by deriving it from a will.'[14]

I first illustrate the attribution of purposiveness to literary works using a classic dispute in literary criticism regarding the 'intentional fallacy.' W. K. Wimsatt and M. C. Beardsley's essay 'The Intentional Fallacy' coins this term to describe their view that literary meaning cannot be understood by searching for authorial evidence of 'intention,' which refers to 'design

[10] Kant 1790, 20:220–1.
[11] Kant 1790, 5:217.
[12] Kant 1790, 5:218.
[13] Kant 1790, 5:222.
[14] Kant 1790 5:220.

or plan in the author's mind.'[15] Rather, they argue, each part of a text 'ought to be judged like any other parts of a composition' in terms of the 'imaginative integration' of each part with the rest.[16] Behind both poles of the debate lies the critical intuition that texts appear 'intentional'—that is to say, willed. In Kantian terms, when we read, the formal qualities of literary works appear to animate our cognitive powers in some purposive way, such that we see in the work a purposive internal causality; we then try to detect that internal causality either by reconstructing the author's plan or by uncovering the structural relationship between the features of the work. As an example, I turn to Eagleton's interpretation of the Marabar Caves in Forster's novel:

> Caves are literally hollow, so that to say that the Marabar Caves lie at the centre of the novel is to say that there is a kind of blank or void at its heart. Like many a modernist work of Forster's time, this one turns on something shadowy and elusive. … If there is indeed a truth at the core of the work, it seems one that is almost impossible to pin down. So the novel's opening sentence serves as a little model of the book as a whole.[17]

Eagleton draws on both authorial intention and textual exegesis to explain the meaning of the Marabar Caves. On the one hand, he attributes the presence of the Marabar Caves to Forster's position as a modernist writer, which made him interested in shadowy and elusive elements. On the other hand, he also relates the image of the Marabar Caves to the overall message of *A Passage to India*. The causal nature of both explanations demonstrates that when we read, we detect a purposiveness in the text that drives us to search for an explanation of what 'caused' its elements to exist and to be placed in a particular order. Therefore, in literary works, 'purposiveness' refers to our sense, as readers, that the formal qualities of the work form a coherent and unified structure.

How, then, does purposiveness in literature arise? I next examine three recent accounts of how the harmonious free play of the understanding and the imagination generates a sense of purposiveness, evaluating each account against the case of literature.

I first turn to Melissa Zinkin's explanation of this activity in terms of extensive and intensive magnitudes. For Zinkin, while 'extensive magnitudes… are measured by homogeneous unitary parts that succeed one another, intensive magnitudes are measured by the degrees contained within them.'[18] So, for example, she argues that length is an extensive magnitude, since 'the length of something… is the sum of the length of its parts.'[19] On the other hand, intensive magnitudes are 'quantitative distinctions among qualities that are the same, such as distinctions between different shades of the same color blue.'[20] For Zinkin, extensive magnitudes are apprehended by the understanding, while intensive magnitudes are apprehended by the imagination.[21] Ultimately, Zinkin accounts for the play of the imagination and the understanding as 'between the multiple apprehensions of a representation by the imagination to form an image of the highest intensity and the understanding seeking to make

---

[15] W. K. Wimsatt and M. C. Beardsley, 'The Intentional Fallacy', *The Sewanee Review* 54, no. 3 (1946): 468–88, 469.
[16] Wimsatt and Beardsley 1946, 484.
[17] Eagleton 2013, 14.
[18] Melissa Zinkin, 'Intensive Magnitudes and the Normativity of Taste', in *Aesthetics and Cognition in Kant's Critical Philosophy*, ed. Rebecca Kukla (Cambridge: Cambridge University Press, 2006), 141.
[19] Zinkin 2006, 140.
[20] Zinkin 2006, 140.
[21] Zinkin 2006, 152.

discursive and uniform this reinforcement of representations.'[22] In other words, Zinkin suggests that the imagination distinguishes between and apprehends the formal qualities of the object to form an intense image, while the understanding homogenises and sums the image produced by the imagination.[23]

Zinkin's account, applied to the case of literature, may mean that the imagination attempts to unite the formal qualities of the text into an overall 'picture' of textual meaning, while the understanding attempts to express this textual meaning as a statement. For example, in Eagleton's reading of the first act of Shakespeare's *Macbeth*, he notices various formal aspects of the work which 'generate an atmosphere of uncertainty, anxiety and paranoid suspicion,'[24] leading to his conclusion that *Macbeth* is about 'human existence, which looks vital and positive enough but is really a kind of nullity.'[25] Applying Zinkin's account to Eagleton, we might say that by noticing these formal qualities, Eagleton's imagination builds up an overall 'picture' of *Macbeth* as a play about suspicion, ambiguity, and confusion, while his understanding uses this 'image' to express a statement about what the play means.

The advantage of Zinkin's account, interpreted in this way, is that it remains compatible with Kant's argument that the play of the imagination and the understanding does not presuppose a determinate concept. By noticing different formal qualities or organising these formal qualities into different statements of textual meaning, critics have offered various interpretations of *Macbeth*. For example, Madelon Gohlke famously read *Macbeth*'s ambiguity as a critique of masculinist violence.[26] Therefore, under Zinkin's account, while the imagination and the understanding work towards mutual agreement on textual meaning, the exact meaning or 'concept' is never determinate.

However, Zinkin's account remains unconvincing when applied to literature. As her example of length in the case of extensive magnitude and colour in the case of intensive magnitude suggest, her account is most readily applicable to the visual arts. It is true that some extensive magnitudes, such as line length in poetry or the temporal and spatial setting of a work, may inform textual meaning, while characterisation in novels might present an example of intensive magnitude in literature—Noël Carroll, for instance, suggests that novels impart moral knowledge by displaying characters who instantiate virtues or vices in varying degrees, which invites readers to make quantitative distinctions between each virtue or vice.[27] However, many literary qualities, such as genre, grammar, or syntax, fail to readily fit into the categories of extensive and intensive magnitudes. Furthermore, Zinkin's account fails to address the issue of purposiveness, or how the interplay of extensive and intensive magnitudes generates a sense of coherent and unified structure in a literary work. Therefore, Zinkin's account fails as an explanation of Kantian aesthetics at least in the case of literature.

As such, I turn to Kenneth Rogerson's argument that formal qualities provoke us towards associations that imaginatively unite into an aesthetic idea beyond empirical

---

[22] Zinkin 2006, 156.
[23] Zinkin 2006, 155–6.
[24] Eagleton 2013, 15.
[25] Eagleton 2013, 15.
[26] Madelon Gohlke, ''I Wooed Thee with My Sword': Shakespeare's Tragic Paradigms', in *Representing Shakespeare: New Psychoanalytic Essays*, ed. Murray M. Schwartz and Coppélia Kahn (Baltimore: Johns Hopkins University Press, 1980), 170–87.
[27] Carroll, Noël, 'The Wheel of Virtue: Art, Literature, and Moral Knowledge', *The Journal of Aesthetics and Art Criticism* 60, no. 1 (2002): 3–26.

experience. He argues that for Kant, artworks 'express ideas of objects or states of affairs beyond our sensible experience by suggesting such things symbolically by way of an analogy.'[28] The symbolic, i.e. formal, elements of the artwork 'stimulate the imagination to make all sorts of associations that substitute for a literal description,'[29] which the understanding then draws together into interpreting that artwork as expressing some non-empirical idea. Rogerson explains that these aesthetic ideas are necessarily non-conceptual as a result of being non-empirical: 'the process of expression is one that must be independent of all 'concepts'—since no concepts can literally describe the notions involved.'[30] The expression of aesthetic ideas, Rogerson argues, then gives rise to a sense of purposiveness, since 'insofar as an aesthetic object can be interpreted as expressing an idea (e.g., creation), it can be seen as exhibiting a kind of organization.'[31]

I illustrate Rogerson's account through G. R. Wilson's reading of John Donne's love-poem, 'A Valediction: of Weeping.' Wilson argues that 'Of Weeping' begins with the image of a tear, but 'as the poem proceeds, the tear becomes, in addition to the coin, womb, medallion, and ocean of the above stanza, the world and finally the entire cosmos.'[32] For Wilson, the transmutational exaggerations of the tear allows Donne to suggest that the lovers' tears are 'the visible manifestation of a far greater spiritual truth—the ideal love that exists in the celestial spheres.'[33] Applying Rogerson's account to Wilson's interpretation, our imagination associatively unites the images offered by Donne's poem, detecting that the tear at the beginning and the globe at the end are linked (both are round) but different (the globe is larger and more significant). Our understanding then assembles this recognition into an expression of an aesthetic idea beyond empirical experience: the poem expresses that the unity of lovers in life prefigures the unity of the self with a loving God. Finally, our feeling that the poem expresses a message generates a sense of purposiveness.

However, as Rogerson himself acknowledges, his account faces two difficulties due to his explanation of purposiveness in terms of expressing ideas. As we will later see, for Kant, purposiveness gives rise to aesthetic pleasure. Therefore, Rogerson's account leads to what he terms the 'everything-is-beautiful' problem: if purposiveness is fulfilled by having an organisation that expresses a non-empirical idea, 'such an account would allow for any object to be aesthetically pleasing since, it seems, any object satisfies this broad aim.'[34] I offer the following example: Ted Sider's paper on mereology, 'Van Inwagen and the Possibility of Gunk,' describes a non-empirical idea—the possibility of a lifeless and atomless 'gunk world.'[35] Since Sider's paper is organised to express an empirically impossible concept, Rogerson's account permits it to be considered as an object of aesthetic pleasure. Rogerson's account thus seems overly permissive insofar as it cannot distinguish between our experience of reading analytic philosophy or literature.

---

[28] Kenneth F. Rogerson, *The Problem of Free Harmony in Kant's Aesthetics* (Albany: SUNY Press, 2008), 29.
[29] Rogerson 2008, 22.
[30] Rogerson 2008, 22.
[31] Rogerson 2008, 92.
[32] G. R. Wilson, Jr., 'The Interplay of Perception and Reflection: Mirror Imagery in Donne's Poetry', *Studies in English Literature, 1500-1900* 9, no. 1 (1969): 107-21, 120.
[33] Wilson 1969, 120.
[34] Rogerson 2008, 67.
[35] Theodore Sider, 'Van Inwagen and the Possibility of Gunk', *Analysis* 53, no. 4 (1993): 285–89.

Secondly, Rogerson's account also leads to the unpalatable conclusion that aesthetic pleasure is rooted in the expression of ideas. He notes:

> If the best interpretation of pleasure in subjective purposiveness turns out to be pleasure in expression of ideas, then it follows that expression of aesthetic ideas is criterial for beauty—a position not widely held by commentators.[36]

Indeed, Rogerson's argument that art expresses aesthetic ideas fails to convincingly account for the case of literature. While literature almost always expresses some cognitive content by virtue of its textual medium, that content may be empirical and our aesthetic pleasure may not derive from the idea expressed. For example, the aesthetic idea behind Keats' 'To Autumn' is that 'autumn is lush and ripe'—an idea that can also be empirically observed and that is so banal that it is unlikely to be the source of our pleasure in the poem. These two difficulties suggest that Rogerson's argument is inadequate as an account of Kantian aesthetics in literature.

While Rogerson addresses these issues by combining his account with a Kantian account of moral judgement, I will deploy an alternative route by combining his account with Malcolm Budd's argument that beauty arises from complex harmony. Budd argues that in Kant's view, the imagination and the understanding serve different functions: 'the imagination feeds on and is nourished by variety and multiplicity (manifoldedness),' while 'the understanding—the faculty of rules—is primed to detect regularity.'[37] In Budd's view, if the perceptual elements are too simple, such as in the case of simple geometric shapes, the imagination grows bored and fails to engage with the work, whereas if the perceptual elements are too diverse, the understanding becomes unable to detect or impose any regularity.[38] Therefore, Budd argues, the artworks we regard as beautiful are complex enough to interest the imagination yet unified enough to facilitate the understanding:

> an object's form will be contemplated with disinterested pleasure when the manifold combined by the imagination is both rich enough to entertain the imagination in its combinatory activity and such as to facilitate the understanding's detection of regularity within it in virtue of composing a harmoniously unified structure.[39]

Budd concludes that purposiveness arises from this sense of structure, as 'elements relate to one another in a harmonious fashion, composing a highly unified whole in which each element appears to be an integral part of the design fittingly related to the other elements.'[40]

I return to Wilson's reading of Donne's 'Of Weeping' to illustrate Budd's account. The poem engages the imagination through its formal complexity of images, from the tear to the coin, the womb, the ocean, and the globe. Simultaneously, the poem facilitates the understanding through its regularity: all the images are round, guiding readers towards its message that the circular unity of lovers is like the unity of the self with God in the celestial

---

[36] Rogerson 2008, 68.
[37] Malcolm Budd, 'The Pure Judgement of Taste as an Aesthetic Reflective Judgement', *British Journal of Aesthetics* 41, no. 3 (2001): 247–60, 258.
[38] Budd 2001, 258.
[39] Budd 2001, 258.
[40] Budd 2001, 258.

spheres. Wilson's analysis exemplifies the sense of purposiveness Budd outlines, as the formal elements of the poem are felt to be constituent interrelated parts of the same design.

Here, I emphasise that Budd's account is especially well-matched to the sense of literary 'purposiveness' I have outlined earlier. When we read literature, we are interested in how the elements relate to the whole work, and our sense of purposiveness arises from the feeling that the elements are joined together in a coherent and unified fashion. Therefore, Budd's account of purposiveness is particularly successful in the case of literature.

I argue that combining Budd's and Rogerson's accounts enriches Budd's account while addressing Rogerson's issues. Budd argues that the imagination delights in the complex variety offered by the formal qualities of the artwork, while the understanding arranges this multiplicity into an overall structure. Rogerson contributes to Budd by explaining *how* imaginative delight in complexity provides the understanding with a sense of structure: for Rogerson, the imagination generates associations stimulated by the formal qualities of the artwork, while the understanding gathers up these associations into the expression of an idea. Simultaneously, Budd's argument that aesthetic experience is simply the detection of complex harmony avoids Rogerson's undesirable implication that the expression of aesthetic ideas is required for beauty; rather, our sense of purposiveness arises from our recognition of a unified structure in the artwork. Furthermore, Budd's emphasis on the limits of perceptual qualities—not too complex but not too simple—ensures that only certain types of organisation give rise to aesthetic pleasure, avoiding Rogerson's 'everything-is-beautiful' problem.

Therefore, I propose a combined Budd-Rogerson account of Kantian aesthetic experience in literature based on 'associative harmony.' In this account, the formal qualities of the artwork stimulate the imagination to generate associations. The greater the complexity and range of the possible associations, the more the imagination is engaged and entertained. These associations contribute to the poem's meaning, which the understanding captures and unites into an overall structure. As readers, detecting the underlying structure behind the appearance of complexity is sufficient to generate our sense that the literary work is purposive, i.e. that it has a coherent and unified design.

Aesthetic Emotion and Judgement: Pattern-Finding as Pleasure and Beauty

For Kant, the sense of purposiveness generated by the harmonious play of the understanding and the imagination arouses a feeling of disinterested pleasure which constitutes our judgement of the artwork as beautiful. As Kant argues, 'the consciousness of the merely formal purposiveness in the play of the cognitive powers of the subject in the case of a representation through which an object is given is the pleasure itself.'[41] For Kant, aesthetic pleasure is disinterested: it neither 'presupposes a need [n]or produces one.'[42] He emphasises that the only desire aesthetic pleasure produces is the desire to continue feeling that pleasure, thus, 'we linger over the consideration of the beautiful because this consideration strengthens and reproduces itself.'[43]

However, Kantian scholars have difficulty explaining how the process of aesthetic judgement remains free of determinate concepts. While Rogerson argues that aesthetic works are non-conceptual by virtue of expressing non-empirical ideas, we have seen that this

[41] Kant 1790, 5:222.
[42] Kant 1790, 5:210.
[43] Kant 1790, 5:222.

argument leads to the unpalatable conclusion that beauty requires the expression of ideas. On the other hand, Budd bites the bullet by concluding that the judgement of beauty requires a concept, namely, the concept of beauty. He argues that the understanding and the imagination deploy 'the concept… of subjective formal purposiveness (that is, beauty!), and… the form is experienced as possessing the property that the concept designates.'[44] In other words, according to Budd, when we recognise the complex harmony in an artwork's form, we apply the concept of 'beauty' to its form to judge that the artwork is beautiful. Given Kant's insistence that the concept presented by the understanding in aesthetic judgement remains undetermined, Budd's conclusion is undesirable.

I demonstrate that the judgement of beauty remains free of determinate concepts by turning to the essay 'What We Mean by Reading' by the literary critic Elaine Auyoung.[45] Auyoung argues that literary critics search for patterns when reading literature—for instance, by looking for repeated words or motifs in a text—because humans find pattern-perceiving inherently enjoyable. She argues: 'researchers of learning and cognition propose that it is pleasurable to discover structures that enable us to perceive a complex set of information in a simpler, more organized way.'[46] Moreover, this aesthetic pleasure is intrinsically pleasurable, since 'according to cognitive perspectives on aesthetic pleasure, attending to literary texts can become interesting and rewarding in itself when we are able to discover in them previously unperceived forms of order.'[47] Auyoung's psychological argument accords with the Kantian account that the perception of purposiveness or 'complex design' underlying a text gives rise to aesthetic pleasure that remains disinterested insofar as it neither presupposes nor produces a need in readers. Rather, as Kant argues, the only desire aesthetic pleasure in literature produces is the desire to linger over the text so that the pleasure strengthens and reproduces itself: in Auyoung's words, 'our interest is sustained by the possibility of perceiving new forms of order that emerge from that complexity.'[48] Finally, Auyoung confirms the Kantian account that disinterested pleasure constitutes the judgement of the work of art as beautiful. She writes, 'the fact that new patterns can still be found in this novel reinforces its status as a complex work of art that rewards further attention.'[49] As my reading of Auyoung shows, associative harmony leading to a sense of purposiveness is enough to create the emotion of disinterested pleasure leading to the judgement of beauty. No determinate concept, be it the expression of ideas or the concept of beauty, is required for aesthetic judgement in literature.

Conclusion

In conclusion, I have offered a Kantian account of aesthetic judgements in literature, while demonstrating that the application of Kant's aesthetics to literature resolves several issues in Kantian scholarship. Firstly, I have shown that literary case studies can adjudicate between competing accounts of what Kant means by the harmonious free play of the understanding and the imagination, enabling me to offer a combined Budd-Rogerson account of associative harmony in literature. Secondly, I have shown that turning to literary scholarship

[44] Budd 2001, 259.
[45] Elaine Auyoung, 'What We Mean by Reading', *New Literary History* 51, no. 1 (2020): 93-114.
[46] Auyoung 2020, 107.
[47] Auyoung 2020, 110.
[48] Auyoung 2020, 107.
[49] Auyoung 2020, 110.

on aesthetic pleasure in literature illuminates Kant's argument that aesthetic pleasure arises from a purposiveness free from determinate concepts. Therefore, I suggest that applying Kantian aesthetics to case studies of specific aesthetic mediums constitutes a fruitful avenue of research for scholars seeking to resolve issues in the *Critique*.

REFERENCES

Auyoung, Elaine. 'What We Mean by Reading'. *New Literary History* 51, no. 1 (2020): 93–114.

Budd, Malcolm. 'The Pure Judgement of Taste as an Aesthetic Reflective Judgement'. *British Journal of Aesthetics* 41, no. 3 (2001): 247–60.

Cheetham, Mark A. *Kant, Art, and Art History: Moments of Discipline*. Cambridge: Cambridge University Press, 2001.

Eagleton, Terry. *How to Read Literature*. London: Yale University Press, 2013.

Gohlke, Madelon. ''I Wooed Thee with My Sword': Shakespeare's Tragic Paradigms'. In *Representing Shakespeare: New Psychoanalytic Essays*, edited by Murray M. Schwartz and Coppélia Kahn, 170–87. Baltimore: Johns Hopkins University Press, 1980.

Kant, Immanuel. *Critique of the Power of Judgement*. Translated by Paul Guyer and Eric Matthews. Cambridge: Cambridge University Press, 2000.

Rogerson, Kenneth F. *The Problem of Free Harmony in Kant's Aesthetics*. Albany: SUNY Press, 2008.

Sider, Theodore. 'Van Inwagen and the Possibility of Gunk'. *Analysis* 53, no. 4 (1993): 285–89.

Wilson, Jr., G. R. 'The Interplay of Perception and Reflection: Mirror Imagery in Donne's Poetry'. *Studies in English Literature, 1500-1900* 9, no. 1 (1969): 107–21.

Wimsatt, W. K., and M. C. Beardsley. 'The Intentional Fallacy'. *The Sewanee Review* 54, no. 3 (1946): 468–88.

Zinkin, Melissa. 'Intensive Magnitudes and the Normativity of Taste'. In *Aesthetics and Cognition in Kant's Critical Philosophy*, edited by Rebecca Kukla, 138–61. Cambridge: Cambridge University Press, 2006.

Time's Transcendental Ideality—Reconciling Kant and His Critics

BENJAMIN OGDEN
McGill University

IN THE YEAR 1770, roughly a decade before he would publish the first edition of the *Critique of Pure Reason*, Immanuel Kant defended his inaugural thesis *Dissertation on the Form and Principles of the Sensible and Intelligible Worlds*. The thesis would propel him to a position that Kant had long coveted—the chair of metaphysics and logic at the University of Königsberg. More importantly for the purposes of this paper, the *Dissertation* contains the early arguments for the application of transcendental ideality to both space and time found in the *Critique*. It is in his inaugural thesis where Kant first famously argues that time constitutes nothing in-itself. In the same year as his defense, Kant sent a copy of the *Dissertation* to his contemporaries Johann Lambert and Moses Mendelssohn. They replied with the most pertinent, and to Kant, concerning, criticisms of his transcendental conception of time. In this paper, I take up Kant's idealistic characterization of time while evaluating his position in the face of the objections proposed by Lambert and Mendelssohn. I proceed to argue that, despite the dissentient debate concerning time's reality in 1770, Kant's conception of time's ideality is not inherently incompatible with his critics' objections. Through analysis of the language employed by Kant, impactful revisions of his earlier ideas on time will be found in the *Critique*. I propose that the subtly reformed position of 1781 demonstrates Kant's attempt to not simply refute his critics, but to cede enough theoretical ground to achieve a degree of reconciliation with his peers.

We begin chronologically with Kant's early work on time in the *Dissertation.* His program foreshadows that of the later *Critique* in that he attempts to demonstrate how space and time are the two fundamental conditions which presuppose and facilitate human cognition. They are enabling conditions, in that, existing prior to any experience, they facilitate our comprehension of spatial extension and temporal succession. Specifically, it is Kant's fifth statement on time that we are primarily concerned with. Here he writes, '*Time is not something objective or real*'.[1] Rather, he claims that time is a 'subjective condition' which is necessary for the mind to coordinate 'all sensible things in accordance with a fixed law'.[2] The claim is that time does not occupy an objective reality but is rather a necessary subjective condition of the human mind. In the same section, Kant will refute any possibility of time occupying an objective reality. Keep this in mind, for we will notice a change to this position by the time the

[1] ID 2:400. Kant, I., Walford, D., & Meerbote, R. (1992). *Theoretical philosophy, 1755-1770* (Ser. The Cambridge edition of the works of Immanuel Kant). Cambridge University Press.; I use this translation for all references to the Inaugural Dissertation.
[2] Ibid.

Critique is published. In Kant's theory, time is a 'necessary' condition because in order to place representations in coordination with principles of succession and change, one must necessarily have an *a priori* intuition of time. The upshot of his claim is that time is not a thing-in-itself, subsisting as a condition generated by human sensibility and persisting as a subjective intuition.

To one familiar with the Transcendental Aesthetic of the *Critique* this argument should not be surprising. Within the Transcendental Aesthetic, Kant explores those pure intuitions which can *a priori* ground representations.[3] Kant does not use 'transcendental' or '*a priori*' in the *Dissertation* to describe time, however, his description of time in the *Dissertation* as being 'presupposed by the senses' reflects the quality of being prior to any experience.[4] Kant will argue in the *Critique* that space and time are the two forms of *a priori* intuition rather than being empirical concepts drawn from experience.[5] This is critical, for if space or time could be learned entities drawn from observation they would not exist as *a priori* enabling conditions. Our observable capacity to gauge temporal succession and spatial extension must necessarily presuppose a pure intuition of time and space. Kant contends that although space is an 'empirical reality', its transcendental ideality entails that space is nothing if we take it up as grounding (external) objects in themselves.[6] That is to say, space is not a thing-in-itself—it is an enabling condition of our senses to coherently organize spatiality.

But these remarks on space are not the source of debate in 1770. In fact, both Lambert and Mendelssohn largely agree with Kant's transcendental conception of space. What is contentious is Kant maintaining that time too is not a thing in-itself. According to Kant, time constitutes a *subjective* condition of human intuition, but outside of the individual subject it is not a thing in-itself, following a similar structure as space.[7] He writes, 'Time is not something that would subsist for itself or attach to things as an objective determination, and thus remain if one abstracted from all subjective conditions of the intuition of them...'.[8] The thought is the following: If one abstracted or removed the conditions which make time intelligible, time would not self-subsist. So, if the objects which I observe to change and express succession were to disappear, time would not remain as an entity, even a non-corporeal one. The pure intuition of time is unintelligible without the initial existence of objects in-themselves. Additionally, the intuition of time would surely disappear if the conscious subject were subtracted. Only once objects in-themselves can be observed to exist can the intuition of time follow and recognize my representations of these objects as being subject to time. If subtracting the interaction between the objects of my representation and my intuition causes the cessation of time, then time cannot logically be a thing in-itself. As in the *Dissertation*, a decade later Kant seems to continue to deny time's objective reality.

With these remarks on Kant's characterization of time, we can take up the criticisms returned to Kant after the circulation of the *Dissertation*. Lambert's and Mendelssohn's responses to Kant take a similar structure: because we can witness change and succession as objectively real, surely, we are committed to time's objective reality as well. In his 1770 letter

---

[3] B34
Kant, I. (1998). *Critique of pure reason*. (P. Guyer & A. W. Wood, Eds.) (Ser. The Cambridge edition of the works of Immanuel Kant). Cambridge University Press.; I use this translation for all references to the 'B' or second edition of the *Critique* which follow.
[4] ID 2:398
[5] B36
[6] B44
[7] B49
[8] Ibid.

to Kant, Lambert proposes this exact thought. He writes, 'all changes are bound to time and inconceivable without time'.[9] After all, change denotes the succession of some object from one state to another over a period of time. On a conceptual level, it is unclear how change could occur without the passage of time. Having established the inextricable link between change and time, Lambert puts forth his next premise that 'even an idealist must grant at least that changes really exist and occur in his representations, for example their beginning and end'.[10] He concludes, 'if changes are real, then time is real, whatever it may be. If time is unreal, then no change can be real.'[11] However, because change can be empirically witnessed as objectively real, Lambert's argument entails that time is real as well. Note that Lambert does not detail what time is, specifically writing of time's reality '...whatever it may be'.[12] Lambert is not arguing for a specific conception of time's ontological composition. His premises appear to be solely concerned with refuting Kant's claim that time is not real.

Mendelssohn's letter to Kant from 1770 mirrors Lambert's argument while proposing a further response. Mendelssohn provides a new proof of succession's objective reality by first taking up the state of the subject's mind, writing, 'Succession is after all at least a necessary condition of the representations that finite minds have'.[13] The representations that a subject might have in his mind can be seen as necessitated to, and bound by, conditions of succession--one thought or mental state can be observed as following another. But Mendelssohn will argue that finite minds must be taken in two states, as both subjects and objects. They are objects in that 'they [finite minds] are also objects of representations, both those of God and those of their fellows.'[14] The thought is not esoteric. Surely an omniscient being like God would know the inner workings of the finite mind and therefore be able to take the mind and its mental states as an object of representation. Obviously, other individuals are not omniscient. However, when two individuals interact with one another, one can infer the mental state of the other based on observation thereby roughly representing their mind as an object. Suppose Jim meets his friend Sarah at a restaurant. In a short period of time, Jim can likely infer the emotional/mental state that Sarah is in at that moment. More importantly, Jim can place these inferences on a temporal chain of succession, from one state to the next. This thought is not definite proof of time's objective reality, but it is evidence of succession's objective reality.

Having shown succession to be objectively real the logical next step for Mendelssohn is to demonstrate that time's reality follows from succession's reality. The argument relies on the relationship between representations and their analogous objects in the world. He writes, 'Since we have to grant the reality of succession in a representing creature and in its alterations, why not also in the sensible objects which are the models and prototypes for representation in the world?'[15] This thought resembles Lambert's and is a compelling criticism of Kant. If the representations in the finite mind are bound to succession, then surely the object being represented is also bound by succession. The representation of an object in my mind undergoing succession or change is objectively real. If we concede that my representation of

---

[9] Johann, Lambert H. *Johann Lambert to Immanuel Kant*, October 13, 1770. Letter. From *Kant Philosophical Correspondence 1759-99.* pp. 63
[10] Ibid.
[11] Ibid.
[12] Ibid.
[13] Moses Mendelssohn. *Moses Mendelssohn to Immanuel Kant*. December 25, 1770. Letter. From *Kant Philosophical Correspondence* 1759-99. pp. 69
[14] Ibid.
[15] Ibid.

succession is only possible through the succession occurring to the actual object itself in the real world, then succession (or change) must be objectively real in that external object. As Lambert has already argued, if succession is an objective reality for the object, then it follows that time is an objective reality as well.

These replies by Lambert and Mendelssohn place the burden of explanatory proof on Kant. The problem Kant faces is that, in the *Dissertation*, he denies that external objects can be taken as being 'in time' or subject to temporality when isolated from the subject. This is his rationale for denying time's objective reality. However, as Lambert aptly puts it, because change and time are bound together, if change can be observed as real then time must also be real. Furthermore, if our representations are subject to and necessitated by conditions of time, the objects of our representations must also be 'in time' and subject to temporality. To argue otherwise would posit an asymmetry between objects and their representation. If my representations are subject to change and succession but the object of my representation is not, then my representation can't be taken as accurately representing the object. It seems that, given these observations, time must have objective reality. Lambert's and Mendelssohn's refutations shift the burden onto Kant to show why time is a subjective condition of human sensibility.

In the *Critique*, Kant does not explicitly respond to Lambert and Mendelssohn, though he considers an objection to time's transcendental ideality proposed by 'insightful men', a likely reference to Lambert and Mendelssohn.[16] The fact that their responses from 1770 remained in Kant's psyche through the publication of the *Critique* suggests their objections concerned Kant tremendously. Following his argument in favor of time's transcendental ideality, Kant lays out the basic objection proposed by Lambert and Mendelssohn a decade earlier: alterations are objectively real, and alterations are possible only with time, therefore time is real.[17] According to Kant, 'There is no difficulty in answering'.[18] His strategy is to concede that time is something real, but that this reality is still truly only subjective. He writes, 'Time is certainly something real, namely the real form of my inner intuition. It therefore has subjective reality in regard to my inner experience'.[19] Note that already this language in the *Critique* of admitting that time possesses some sort of reality represents a shift from the rigidness of the *Dissertation*. Kant is not just repeating his dissertation verbatim. However, in order to get at the crux of Kant's strategy in this response and to understand time's supposed subjective ideality yet objective reality, we must briefly explore what Kant means by one's 'inner experience' or inner sense.

Earlier in the Transcendental Aesthetic, Kant distinguishes between two senses, the inner sense and the outer sense. Time and space, the two forms of pure *a priori* intuitions, correspond to these respectively. In his second concluding remark on space Kant writes, 'Space is nothing other than merely the form of all appearances of outer sense'.[20] Like time, space is also a subjective condition of sensibility 'under which alone outer intuition is possible'.[21] The pure intuition of space is a condition of the outer sense because it intuits appearances in the spatial world external to the subject. Kant writes that 'if we depart from the subjective conditions, under which alone we can acquire outer intuition...then the representation of space

[16] B53
[17] Ibid.
[18] Ibid.
[19] Ibid.
[20] B42
[21] Ibid.

means nothing'.[22] The claims is that if we were to abandon the perspective of the subject, no representation of space could be given. Further entrenching the distinction between inner and outer sense is the observation that, because physical spatial extension does not exist within the mind, the intuition of space concerns itself with the outer sense alone. In his second concluding remark on time, Kant distinguishes time from space, writing, 'Time is nothing other than the form of inner sense, i.e., of the intuition of our self and our inner state. For time cannot be a determination of outer appearances, it belongs neither to shape or a position etc. but on the contrary determines the representations in our inner state'.[23] Kant's commitment to time's subjective reality relies on his understanding of time as being inextricably linked to the inner sense.

Earlier I proposed that we can conceptualize these intuitions of space and time as enabling conditions. This is supported by Kant's exposition on the distinction between the inner and outer sense. With the intuition of space, the capacity for objects to interact with a subject is enabled through a prior intuition of spatial extension in the subject, explaining why this is the form of the outer sense i.e. external objects. However, 'if we depart from the subjective conditions, under which alone we can acquire outer intuition...then the representation of space means nothing'.[24] Abandoning the first-person subjective entails that a representation of space becomes impossible precisely because it is the intuition generated in the mind which enables objects to be ordered in spatial extension. Conversely, time 'cannot be a determination of outer appearances, it belongs neither to shape or a position etc. but on the contrary determines the representations of the inner state.[25] Yet similarly to space, Kant holds that abandoning the first-person subjective makes time unintelligible as its own entity. But the prime question under investigation is whether this contention that time is a subjective enabling condition entails that time is not something real.

The intuition of the inner sense, and therefore the intuition of time, is the intuition of only two things: our self and our inner state. Whether we can apply the concept of self-consciousness to Kant's conception of the inner sense is unclear. In his exposition of space Kant does write that 'Inner sense, by means of which the mind intuits itself, or its inner state, gives, to be sure, no intuition of the soul itself as an object...'.[26] So, while the inner sense does not necessarily take myself or my soul as an object, there is a level of self-reflection occurring when the mind intuits itself. So, suppose I am angry one moment and content the next. Though my attitudes do not have spatial embodiment, the intuition of time allows the mind to judge its inner state (and the mental states it forms) as passing from one attitudinal state to another. More importantly for this paper however is how Kant takes representations to relate to the inner sense. Kant writes, 'whether or not they have an outer thing as their object' representations are determinations of the mind and therefore of the inner state.[27] So, whether or not a representation corresponds to an appearance that is intuited by space, the representation belongs to the inner state. Because time is the formal condition of the inner state, all representations exist under the

[22] Ibid.
[23] B49
[24] B42
[25] B49
[26] B37
[27] B50

intuition of time—'everything that belongs to inner determinations is represented in relation to time'.[28]

These remarks leave no doubt that Kant takes representations to be subject to conditions of time, but they also assist us in understanding how Kant might work through the objections of Lambert and Mendelssohn. In what we have explored so far, Kant has always dealt with appearances and representations, never objects in-themselves. Because our representations of objects are the products of our sensibility, they are subject to the intuitions of space and time. But Kant notes that if we 'take objects as they may be in themselves, then time is nothing'.[29] As appearances are objects of our sense, time does have an 'objective validity' from the subjective perspective, but if one takes an object as abstracted from our sensibility time loses all objective reality.[30]

Kant therefore establishes an important distinction between the world of our sensibility and the world removed from the human subject where objects are taken as things in-themselves. If time is a formal condition of the inner sense and one subtracts that inner sense from all objects and therefore all appearances, the objects in-themselves would not be under any condition of time. Kant's position seems to contend that the condition of time is *generated* by the subject. This does not strike me as too dramatic a description of Kant's stance. After all, Kant writes himself that time does not occupy any shape or position and cannot be intuited externally.[31] Rather time is a condition bound to the inner sense therefore seated in the mind of the subject. As I have been suggesting, Kant's account in the *Critique* therefore takes a first-person perspective. The *subject's* presence in the world determines whether objects are 'in time' in the form of appearances and representations or alternatively not 'in time' as things-in-themselves. Recall the common contention of Lambert and Mendelssohn that if changes are real then time must also be real. More fully put, the argument is that if our representations can be intuited as changing then why shouldn't the objects themselves that are being represented also be taken as changing? If we concede this point then it follows that if our representations are subject to the condition of time, then the objects themselves are also subject to time or can be said to be 'in time'.

Inherent in the responses of Lambert and Mendelssohn is an assumption of the objective reality of objects in-themselves. Bracketing our discussion of time for a moment, Lambert's and Mendelssohn's arguments possess a crucial assumption that the reality of objects in-themselves is undeniable. Kant's capacity to refute Lambert and Mendelssohn (or at least to resolve the debate) is partially contingent on raising cause for skepticism about the validity of this premise. If Kant can question our capacity to engage with objects in-themselves, he can dispel with Lambert's and Mendelssohn's' objections which rely on the objective reality of objects in-themselves. It is in this stratagem that we find the potential for reconciling the broader debate concerning time—Kant does not seem to deny that our representations are in time rather he denies that time is a thing in-itself outside of the realm of my representations and therefore outside of the first-person subjective perspective.

Kant asserts that, 'the reality of outer objects is not capable of any strict proof.[32] On the contrary, he argues that it is the reality of the object of our inner sense which is 'immediately

[28] B37
[29] B51
[30] Ibid.
[31] B49/B37
[32] B55

clear through consciousness'. This would be a controversial position to Lambert and Mendelssohn because it inverts their initial assumption. Neither Lambert or Mendelssohn have a problem with admitting the reality of outer objects in the world. In fact, there position positively argues that changes and time in outer objects are both objectively real. But Kant is skeptical of the reality of outer objects. He is not denying their existence, but he is questioning our capacity to provide a *proof* of their existence. This skepticism is grounded in the claim that both our representations in the inner sense (our mind) and the objects of our outer sense 'belong only to appearance' not to objects in-themselves.[33] By questioning our capacity to engage with object in-themselves, Kant can demonstrate that the boundary of time is limited to the realm of sensibility (representations and appearances) 'beyond which no further objective use of them [space and time] takes place'.[34] Therefore, the fact that the *a priori* intuition of time is limited to appearances entails that it cannot, by its very nature, be objectively real beyond the reach of sensibility to objects in-themselves. If I can only intuit objects by way of appearances, time is forever limited to appearances and cannot be subject to objectively valid proof beyond the conditions of sensibility.

This response places some of the explanatory burden back onto Lambert and Mendelssohn. After all, in their letters to Kant they do not provide a formal proof of the reality of objects in-themselves. It is simply assumed. However, Kant's response has its own limitations. Lambert and Mendelssohn could simply reply that Kant is engaging in a degree of circular reasoning. The only reason Kant can claim that time is bound to appearances and not objects in-themselves is because he has already defined time as a pure intuition that constitutes a subjective condition of the inner sense and as a product of sensibility. Time is limited from objects in-themselves because of Kant's prior definition of the limits of sensibility. Kant's argument that time is a subjective condition relies on time being bound to the sensibility as a pure intuition, but his characterization of time as this sort of intuition relies on time being a subjective condition of the inner sense. So, while Kant's response is certainly addressing the relevant criticisms, he doesn't seem to possess the theoretical content to demonstrate time's subjectivity without collapsing back on his own theory of time as a pure intuition.

Despite the limitations of Kant's response to his contemporaries, I propose that the language of the response leaves the door open for reconciliation. Paul Guyer's recent discussion of this subject argues that Kant's defense of time's ideality in the *Critique* would be insufficient to Lambert and Mendelssohn.[35] After all, as Guyer points out, much of the response found in the *Critique* is simply reiterating Kant's initial argument that time is nothing more than an intuition of the inner sense—an enabling condition which makes observable change and succession intelligible.[36] This would be seemingly incongruent with the letters of both Lambert and Mendelssohn. However, to avoid 'talking past' our three interlocuters, let us not lost sight of what sparked this debate in the first place. The discourse between Kant and his critics was generated by Kant's denial that time is nothing in-itself. Lambert and Mendelssohn both respond directly to this claim, reflected in their arguments in favor of time as being something positively real. Recall for instance the argument that if change can be observed as real, time

[33] Ibid.
[34] B56
[35] Guyer, P. (2020). Reason and experience in Mendelssohn and Kant (First). Oxford University Press. pg. 182. Unfortunately, Lambert and Mendelssohn never directly respond to Kant's refutation of their argument in the *Crtique*.
[36] Ibid.

must be something real as well. The argument is designed to demonstrate time's objective reality and dispel any chance of its transcendental ideality. But, by the time of the *Critique,* when does Kant explicitly deny time's reality?

In his letter from 1770, Lambert writes that he finds Kant's first four statements on time in the *Dissertation* 'quite correct'.[37] Lambert has no qualms with Kant's claims that time is presupposed by the senses or that it is a pure intuition etc. The point of contention is solely that Lambert does not believe time can be 'regarded as something unreal'.[38] So, the only relevant question at this stage of inquiry is whether Kant does actually claim that time is something unreal. Admittedly, as we have seen, he does so in the *Dissertation*. But, if we view the *Critique* as a partial response to the dialogue of 1770 then it is within the *Critique* that we should look for Kant's most developed position on the matter of time. By 1781 (the year that the first edition of the *Critique* was published), Kant concedes time's *empirical reality* as a subjective condition of representation.[39] But if we concede time's empirical reality, how can time not be a thing in-itself? If time can be empirically observed, Kant is seemingly committed to time's objective reality as an empirically observable phenomenon. Before working out this problem, note that thus far Kant's position is in line with Lambert's and Mendelssohn's. The argument whereby time's reality is deduced from the observable (and therefore empirical) reality of time is mirrored in Kant's concession of time's empirical reality.

CONCLUSION

The question is therefore not a matter of time's reality. Rather, we have been unknowingly dealing with a subtler debate concerned with the *nature* of time's reality. If Lambert and Mendelssohn argue that time is objectively real, it isn't apparent that Kant disputes this claim *prima facie*. Indeed, by the publication of the *Critique,* it is evident that Kant's ideas concerning time have evolved to appease the criticisms of 1770. Where, in the *Dissertation,* Kant is quick to deny time any objectivity whatsoever, a decade later in the *Critique* Kant has toned down his language to allow room for time to possess a modicum of objectivity. Here Kant explicitly writes that the empirical reality of time is of 'objective validity in regard to all objects that may ever be given to our senses'.[40] The crucial modifier here is that time's objective validity is contingent upon the relevant objects, observed to be 'in time', being supplied to our senses. The point of potential dispute only comes when Kant denies time's absolute reality, that is the claim that time could persist as a thing in-itself when separated from our sensible intuition of the objects considered to be within time.[41] Kant's position, as I have noted, is that of a skeptic. Suppose he is correct that my knowledge of objects in-themselves is forever limited to my interaction with them as appearances and representations by way of my intuition. If I cannot have any knowledge of entities as objects purely in-themselves then surely, I cannot positively know whether these objects in-themselves are in time. My knowledge of objects and their accessibility to me is forever limited to the realm of sensibility. My intuition

[37] Johann, Lambert H. *Johann Lambert to Immanuel Kant,* October 13, 1770. Letter. From *Kant Philosophical Correspondence 1759-99.* pp. 62. Mendelssohn's letter is much shorter, so it is more difficult to know how much he agreed with the *Dissertation*.
[38] Johann, Lambert H. *Johann Lambert to Immanuel Kant,* October 13, 1770. Letter. From *Kant Philosophical Correspondence 1759-99.* pp. 63
[39] B52
[40] Ibid.
[41] Ibid.

of time is therefore tied to my sensibility and classified as a pure intuition, not because time necessarily lacks objective reality, but because I cannot have knowledge, positive or negative, of objects in-themselves being in time.

By 1781, we can witness ground being ceded to Lambert and Mendelssohn. But note that Kant will only cede as much as necessary. He will admit time's objectivity while denying the absolute scope of this objectivity in order to remain faithful to his overall program. In doing so, Kant walks a thin line between conceding time's objective reality while arguing that this objective reality is only objective, or empirical, within the scope of a the first-person subjective. That is to say that if one removes the conditions of sensibility generated in the first-person subjective, then we lose all knowledge of the object in-themselves and therefore cannot speak of time in any coherent way. Time is of 'objective validity' regarding appearances and representations because, through one's sensibility, one comes to have knowledge of the object being represented.[42] Again, and again, throughout his conclusions on time in the *Critique*, Kant reaffirms time's objectivity within the realm of what I have been calling the first-person subjective. This conception of time as being objectively real within the first-person subjective may appear oxymoronic. However, we must remember that Kant deals deliberately within two realities—the reality of objects in-themselves and the reality of objects as the appearances and representations of our sensibility. What is true within one realm does not necessarily hold in the other. So, while time's reality may lack objective proof in the sphere of objects in themselves, if we speak of appearances and representations then time's reality may very well be capable of objective proof. Far from grudgingly admitting time's reality, Kant positively argues for time's objectively real nature insofar as we consider objects within our sensibility.

But we must not forget that, despite its nuanced presentation, this admission is a major leap from Kant's position in the *Dissertation*. Kant's recognition of Lambert's and Mendelssohn's criticisms as substantive arguments encourages our understanding of the Transcendental Aesthetic as Kant's attempt to rework his theory in light of 1770. The discourse I have developed thus far must not be viewed as two distinct episodes. The debate of 1770 and Kant's conception of time's reality in the *Critique* ought to be taken as a single unified dialogue. Were one to read the dialogue of 1770 alone, one would conclude that Kant and his critics remain forever opposed on the topic of time's reality. It is only through analyzing 1770 and 1781 together, which has been the project of this paper, that we see Kant's updated position as potentially reconcilable with Lambert and Mendelssohn. The revisions Kant makes to his own theory of time's transcendentality in the *Critique* represent the final word in the debate. The episode of 1770 through the *Critique* is therefore as much a history of Kant's thoughts on time as it is a debate with his peers. And, in the *Critique*, we find a clear attempt by Kant to reform his theory considering the criticisms from 1770. We cannot know how Lambert and Mendelssohn would have responded to my comments in this paper nor to Kant's response in the *Critique*, but this paper's object has been to reorient how the exchange from 1770 and Kant's further writings on time in the *Critique* should be viewed.[43] The section on time in the *Critique* is not simply a pure rebuttal of Lambert and Mendelssohn while restating the position of the *Dissertation*. It represents a revised model of Kant's theory which should be taken, not as

[42] B51

[43] By the time that Kant published the first edition of the *Critique of Pure Reason* in 1781 Johann Lambert had already died in 1777. Moses Mendelssohn would live to see the publication of the *Critique* but the infirmaries of his old age, noted by himself in his letters to Kant, prevented him from engaging with the *Critique* on a critical level.

simply a refutation of the 1770 criticisms, but as an attempt to reconcile the discordant dialogue of 1770.

REFERENCES

Guyer, Paul. (2020). Reason and experience in Mendelssohn and Kant (First). Oxford University Press.

Lambert, Johann. (1967). *Johann Lambert to Immanuel Kant*. October 13, 1770. Letter. From *Kant Philosophical Correspondence 1759-99.* University of Chicago Press.

Kant, Immanuel, Walford, David, & Meerbote, Ralf (1992). *Theoretical philosophy, 1755-1770* (Ser. The Cambridge edition of the works of Immanuel Kant). Cambridge University Press.

Kant, Immanuel. (1998). *Critique of pure reason*. (Paul Guyer & Allen W. Wood, Eds.) (Ser. The Cambridge edition of the works of Immanuel Kant). Cambridge University Press.

Mendelssohn, Moses. (1967). *Moses Mendelssohn to Immanuel Kant*. December 25, 1770. Letter. From *Kant Philosophical Correspondence 1759-99*. University of Chicago Press.

Dormitive Virtues Are Not a Weakness: A Defense of Powers

EUGENE TAKEUCHI WILLIAMS
Durham University

THIS ESSAY is a brief defense of causal powers from a popular and persistent foe. 'Virtus Dormitiva' argues to the effect that powers are an ineffective explanation of causation, because they provide no new information about the causal scenario. I assume a basic familiarity with powers, as well as Humean ontologies.

Ultimately, I believe it can be stated confidently that Virtus Dormitiva fails. Its failure comes in many possible forms: it's possible that the argument fails to address its opponent, because the analogy on which it rests is a rather uncharitable one. On the other hand, Virtus Dormitiva may fail to show exactly what the issue is with an appeal to 'dormitive virtues'. Finally, I raise suspicions that the argument is merely an empty vehicle for long-held prejudices against non-Humean ontologies, since it is unclear that the problems highlighted by Virtus Dormitiva are unique to Power Theory.

The Virtus Dormitiva (VD) argument is originally taken from Molière. Neil Williams paraphrases the story on which it is based:

> a doctoral candidate undergoes oral examination by a group of learned doctors to determine if he is worthy of joining their ranks. When asked by the doctors for the cause and reason that opium makes one sleep, he replies that it possesses a *virtus dormitive*: that is, a power whose nature is to cause sleep. The doctors are most pleased with his response, and he is sworn in.[1]

The story mocks powers as a genuine causal explanation. Indeed, there seems to be something tautological about the explanation the doctoral candidate employs, which we can immediately detect. But if it is to be an argument against power theories (say, based on dispositional properties and manifestation processes), the argument needs refining. There are two tests which VD needs to pass to be effective:

(A) VD must demonstrate a parallel between the doctoral candidate's reasoning and powers explanations.

(B) Given the parallel, VD must clarify the weakness of an explanation based on powers.

[1] Williams, Neil, *The Powers Metaphysic*, p. 196.

If one can show convincingly that power theory commits the same errors as the doctoral candidate, and that those errors are indeed fatal, a powers-based explanation of causation cannot be desirable, even if it is plausible *prima facie*.

Firstly, (A). Does a sophisticated power theory really say what VD claims it does? If not, VD fails to address its opponent. Obviously, on the surface of things, it would be extremely uncharitable to argue for an exact parity of reasoning between the would-be doctor and power theory. Where are the manifestation processes? Where is the argument against Humean views? Where are the irreducible dispositional properties? But VD can still argue that, despite all the details of any specific power theory, the causal *explanation* which is given by powers will never change, at least at its core. Under power theories, the answer to 'what caused M?' will always be 'something which was disposed to M'. So, the argument is that powers-based explanations can be reduced to this. VD argues that powers are 'vacuous causal explainers',[2] and even sophisticated power ontologies fail to explain causal phenomena in a non-trivial way.

Presumably, one would go about proving this is by taking a powers-based explanation which is, on the surface, explanatorily *non*-vacuous, and show that it too appeals to the same reasoning. Let's try, with an example taken from Cartwright and Pemberton. It concerns a powers-based description of a toilet and its flushing mechanism:

> When the handle (flush control) is turned, the lever arm (which has the power to pull up the lift rod) pulls up the lift rod (which has the power to open the outlet valve) which opens the outlet valve (which has the power to release water from the cistern), which releases water from the cistern, which lowers the ballcock, which opens the inlet valve.[3]

This example from engineering is a good example of power theory's strength, in that it demonstrates the conceptual relevance of powers to typical causal scenarios. It doesn't seem trivial or meaningless to give the above as an answer to 'how does the toilet refill with water after one flushes it?'. But if powers are vacuous causal explainers, this explanation too, must be vacuous. Presumably for those who argue for VD, the problem is that power theory also allows the macro-level power of the toilet 'to refill itself when flushed'. Any answer to the previous question which refers to such a power once again seems vacuous, in the same way Molière wished to highlight. But these are two different explanations based on different powers, and power theorists will be the first to admit that the power 'to refill itself when flushed' is far too macro in scale for the question 'how does the toilet refill?'. For instance, Williams recognises that 'it is not uncommon to arrive at power ascriptions via a process of reverse engineering, working backwards from the behaviour to be explained. [In which case,] no new information about the behaviour [is given], and therefore, […] the postulation of the power is uninformative.'[4]

It's hard to say definitively whether the existence of the uninformative, macro-level powers is a weakness of power theory, when in those cases there will also be lower-level powers more suited to causal explanation. But I would argue this is not a weakness for power theories, because causal explanation, and the implicit criteria for effective explanations (whatever they may be exactly) are not a concern for the power ontology, but rather for the

---

[2] Mumford, Stephen and Anjum, Rani Lill, *Getting Causes from Powers*, p. 133.
[3] Cartwright, Nancy and Pemberton, John, M., *Aristotelian Powers: Without Them, What Would Modern Science Do?*, p. 100.
[4] Williams, Neil, *The Powers Metaphysic*, p. 197.

epistemology of science. In other words, the fact that there are macro and micro-level powers needn't tell us anything about *explanations* based on them. Power theory – by which we mean a species of metaphysical and ontological practice – doesn't tell us anything about how science ought to discover powers, and it also doesn't tell us how science should use powers in their explanations, or what good explanations are in general. This seems like a strength, because the criteria for a 'good explanation' are notoriously hard to pin down across the sciences. So, it seems as though VD's criticism should be directed at the potentially bad explanations based on inappropriately chosen powers, rather than powers themselves.

One last effort to try to get VD to pass test (A). Although it's true that power ontologies don't tell us directly about how powers ought to be used in explanations of causation, it's still possible that the world bottoms out at fundamental powers. Many power theorists believe this: Dispositional Monists will be *forced* to commit to this view, because they suppose everything that exists *is* a power. Mixed View advocates may believe that *in some cases,* fundamental powers will have to be referred to, since not everything has a categorical base. If so, we can follow a chain of "Why?" questions until we get to a hopelessly Molièresque answer, as follows:

> "Why does the toilet refill with water after one flushes it?"
>
> "Because a series of powers, starting with the ones possessed by the flush handle, releases water from the cistern"
>
> "Why does the flush handle release water from the cistern?"
>
> "Because the handle has the power to pivot on its axis, pulling the lift rod with it"
>
> "Why does that clump of molecules have the power to pivot rather than snap?"
>
> "Because of the handle's structure – the molecules which compose it are packed tightly (this can, with more effort, be rewritten in purely dispositional language, although my language here is more akin to a Mixed View, where molecular structure is a categorical property)"
>
> "Why are the molecules packed tightly?"

At this point we are forced to talk particle physics, exchanging fundamentally dispositional terms like mass, charge and spin, as well as forces which are brought about by virtual particles. Presumably, since there are no more novel powers to refer to, when we ask why a certain particle or collection of particles have a certain charge or spin, we can only answer as the doctoral candidate does: the particles can *do* such and such, because they have the *power* to do such and such. Perhaps this *is* uninformative.

But, I would reply, irreducible powers were always a part of the theory. Being left with irreducible powers at the most fundamental level doesn't come as a surprise to the power ontologist. So perhaps the more honest way to counter VD is to argue against (B); there is nothing *necessarily* wrong with appealing to powers in a circular way.

Williams does this by appealing to a certain 'ontological relativity of explanation':[5]

> Neo-Humean ontologies countenance laws and categorical properties but treat powers as second-class. Powers are therefore not proper parts of causes, and so do not figure in good explanations. But this […] cuts both ways. If the correct ontology grants powers full status as causal properties (as the powers ontology says it does), it follows […] that powers-based explanations that reference them would be good explanations.[6]

In other words, if we are supposed to be suspicious of powers-based explanation simply because they employ powers, the problem is with the questioner. If one 'switches in' an ontology which is more accepting of powers, and finds that the challenge posed by VD dissipates, the problem only existed *between* ontologies, not *within* the powers ontology.

This is a convincing argument, because it is true that, from a power-friendly perspective, there is nothing out of the ordinary about the description of the toilet handle in terms of the fundamental powers of its constituent particles. There is also nothing wrong with saying that those particles behave the way they do because of fundamentally dispositional properties. There is nothing wrong with this explanation from a scientific perspective because physicists are investigating how these particles behave, and assign properties based on experiments. There is nothing wrong with it at the level of powers either, because, having reached the most fundamental properties, there is no lower level to appeal to, and irreducibility is one of the key features of powers at the fundamental level.

One could follow this relativistic point through to its conclusion. A Humean ontology, only allowing categorical properties (those which do *not* dispose the object to do anything) would reach the same circularity of explanation. Suppose a power is a second-class property which is actually reducible to a conditional analysis – the details needn't worry us, but suppose it's true. Then, powers are at bottom just categorical properties related through conditionals about 'What would happen to object O, if x occurred?'. Following the same line of inquiry as before, we'd still be left with circularity at the fundamental level. Nothing more can be said about the most fundamental categorical properties, than that 'they are how they are'. Why are the molecules packed tightly? A Lewisian would say 'because the laws of nature dictate how the particles interact'. But Lewisian laws are just generalizations of what happens to be true in a given world. So, we'd be left with unexplainable matters of fact – the particles are the way they are. And scientific causation would be based on these matter-of-fact laws of nature.

If Humeans and power theorists alike will reach the bottom of the well at some point, it makes even more sense to double down on powers: *there is nothing wrong with appealing to irreducible powers at the fundamental level*. If there is anything wrong with how Molière's doctoral candidate explained opium's power to induce sleep, it is that he got the science wrong; opium's dormitive virtue is in fact shorthand for a number of lower-level powers. But even those of us who like to think we 'get the science right' will regularly make reference to 'somniferous'[7] medicine and food, which signifies nothing more than that X possesses a dormitive virtue.

---

[5] Williams, Neil, p. 199.
[6] Ibid.
[7] Michon, Cyrille, *Opium's Virtus Dormitiva*, p. 134.

REFERENCES

Cartwright, Nancy and Pemberton, John, M. 'Aristotelian Powers: Without Them, What Would Modern Science Do?' *Powers and Capacities in Philosophy: The New Aristotelianism*, 2013, 93–112.

Michon, Cyrille. 'Opium's Virtus Dormitiva'. In *Dispositions and Causal Powers*, 133–50. Ashgate, 2007.

Mumford, Stephen and Anjum, Rani Lill. *Getting Causes from Powers*. Oxford University Press, 2011.

Williams, Neil. *The Powers Metaphysic*. Oxford University Press, 2019.

Epistemic Justification and the Interpersonal View of Testimony: The Epistemic Power of Passing-the-Buck

LAUREN SOMERS
Cambridge University

TESTIMONY IS A PECULIAR and much-discussed source of knowledge. On the one hand, many of our epistemic institutions are founded on the idea that testimony can in fact grant epistemic justification to a belief. Its doing so is extremely valuable, facilitating the generation of collective knowledge by societies, and making knowledge generally easier to come by. On the other hand, what kind of epistemic justification testimony grants is largely unclear. As demonstrated by arguments made by proponents of the interpersonal views of testimony (IVTs), it is not obvious that the justification granted by testimony is of the same kind as the justification that is granted by other modes of epistemic justification. There are key differences in both the intrinsic and the extrinsic features of testimonial justification which distinguish it, suggesting that it grants a *sui generis* species of testimonial warrant. However, if the epistemic justification granted by testimony is of its own kind, distinct from other forms of epistemic justification, the question has been raised of how it is epistemic at all. Since the debate is normally couched in terms of evidence, I shall be focussing on the contrast between testimony and evidence, and IVT and evidentialism. Many of the points, however, carry across, with appropriate substitutions, to other epistemic views.

In this essay, I explore the above question through a close examination of IVT and its criticisms. In particular, I consider how the interpersonal theorist might respond to the objection that their account renders testimony epistemically impotent, and establish where within the general landscape of epistemic justification they might best locate themselves to refute this claim. The problem with the view is in explaining how interpersonal features can be epistemically relevant in a distinctive, non-reductive way. I argue that Jennifer Lackey, who makes this complaint, fails to take account of the great variety of theories of epistemic justification which the interpersonal theorist might associate their theory with.

Though IVT highlights features of testimony which differentiate testimony and evidence, I suggest that there is a suitable revision in how we construe epistemic justification which makes the interpersonal view *compatible* with some version of evidentialism. Testimony and evidence are not the same, but are connected in that testimony presents a unique, *sui generis*, distinctively social way of being connected to some further item of evidence. This differentiates testimony from other sources of epistemic justification: it is *only* in the case of testimony that we can garner epistemic justification for a belief that *p* without the source having to be evidence for *p*.[1] This revision has consequences for how we view epistemic justification

[1] As McMyler (2011) argues, we might take the fact that we can treat testimony as evidence, and also treat it as *sui generis*, and use this to endorse some kind of pluralism. This, however, is undesirable. We

more generally: the best account of epistemic justification, once we have taken into account the interpersonal theorist's observations, makes the root of epistemic justification the fulfilment of epistemic duty. The revision makes room for our social relationships to bear epistemic weight in a way which comes apart from merely treating the people in the relationships as truth-gauges, indicators of the likelihood of the belief.

There are three steps we must take to establish this paper's thesis. First, we must establish the common ground. This consists in defining the key ideas, describing the features that differentiate testimonial and evidential justification, and setting out Lackey's dilemma with the problem it poses. Next, we must show how the interpersonal theorist can counter this dilemma. There are two steps to this. Step one is to establish by what mechanism exactly testimony generates testimonial knowledge. Step two is to establish precisely how this warrant can be epistemic – how the interpersonal view fits into the epistemic territory, and how in fact the interpersonal view of testimony can be consistent with evidentialism. Last, and to conclude, we must evaluate the implications of this response to Lackey's dilemma on our theories of epistemic justification more generally.

PART 1: TESTIMONIAL WARRANT IS NOT EVIDENTIAL

1. DEFINITIONS AND IVT

To start off with, we say that a belief, *p*, is testimonially derived iff someone tells us a statement expressing *p*, and we accept it. This is not restricted to testimony in the formal sense of courtroom testimony,[2] rather we are interested in 'tellings generally with "no restrictions either on subject matter, or on the speaker's epistemic relation to it"'.[3] This understanding of testimony leaves open the question of whether the speaker must themselves *know* that which they are testifying to, or even whether they must have formed their belief about it in an epistemically responsible manner. On such questions I shall remain as impartial as possible. As Lackey does, I take the crucial notion in this definition to be that of 'telling', which entails that there be an intention on the part of the speaker to convey information for a speech act to count as testimony.[4] This rules out cases of sleep talk, for example, or recitations of poetry, as cases of testimony: they are speech acts, but they are not intended to convey information (or at least not straightforwardly so). In Lackey's formulation, testimony is 'a speaker's making an act of communication – which includes statements, nods, pointing, and so on – that is *intended to convey the information that p.*'[5]

---

would end up holding both that the act of testifying grants its own kind of warrant, which is not evidential warrant, and that it grants evidential warrant. Lexicographically, this is confusing, and not an awful lot is gained by thinking in this way. It seems better overall to think that, if it's evidential it's not strictly testimonial, but you could treat testimony as evidence, if you so wished. This maintains testimonial warrant as a unified kind.

[2] Interestingly, many traditional accounts of testimony, particularly non-reductive and interpersonal accounts, would rule out courtroom statements as 'testimony' in a true sense. This is because the statements in question are not freely given, as they are taken under oath.

[3] Fricker 1995 in Lackey 2010, 2.

[4] Moran also accepts an intention condition, though such a condition is sometimes disputed (Adler 2017).

[5] Lackey 2010, 3. Emphasis mine.

Further, it is distinctive of testimonially justified belief that the warrant is grounded in the *content* of a speaker's testimony'.[6] So Richard Moran says we want to 'account for what is distinctive about acquiring beliefs from what people say, as opposed to learning from other expressive or revealing behaviour of theirs'.[7] This rules out cases where, for example, you learn that someone has spoken on the basis of a person's testimony, as cases of genuinely testimonially justified belief.[8] It does not, however, rule out that the testimonially justified belief might be combined with some other justifications on its pathway to knowledge.

We can sort theories of the epistemic powers of testimony into two categories: reductive and non-reductive. Reductive views entail that the epistemic justification acquired by testimony is reducible to justification acquired by other epistemic faculties, such as perception, memory, and induction; non-reductive views entail the contrary.[9] IVTs are in the non-reductive category. Interpersonal theorists hold that the special justification acquired by testimony is a property in particular of the interpersonal relationship between the testifier and the recipient of testimony. They are committed to the claim that the epistemic value conferred upon beliefs acquired through testimony is granted by features of this interpersonal relationship.[10] The relationship between participants is directly responsible for the capacity of testimony to generate or transfer knowledge. The two most prominent traditional versions of IVT are Moran's 'Assurance View' on which an epistemic warrant is generated by the testifiers offering an assurance of the truth of the belief, and the recipient accepting it,[11] and Edward Hinchman's 'Trust View', on which an epistemic warrant is generated by the testifier offering an invitation to trust them, and the recipient accepting it.[12]

There is, however, a third thesis which it has been suggested the interpersonal theorist must hold: that the epistemic justification which is granted by the interpersonal features of the relationship must be '*non-evidential* in nature'.[13] We shall come to the justification of why the interpersonal theorist might think this in the next section, but for now it's worth getting clearer on what evidence is, and what might constitute evidential justification. The nature of evidence is its own thorny issue, and I shall try to side-step most of the specifics, since much of what I shall say is compatible with many different conceptions. For the purposes of this paper, I adopt only the minimal condition that evidence for a hypothesis must at least make the hypothesis more likely.[14] This is neutral between internalist and externalist accounts of evidence, as it is neutral between subjective and objective probability. It is also neutral on some other questions in the area, notably on what sorts of things might be evidence, whether they be propositions or events. Then, we might say that a belief, *h*, had by *S*, is evidentially justified just in case the evidence, *e*, is believed by *S*, and *e* makes *h* more likely.[15]

---

[6] Ibid.

[7] Moran 2005, 3.

[8] I exclude the case where the speaker is testifying that they have spoken.

[9] Green 2020.

[10] Ibid.

[11] Moran 2005.

[12] Hinchman 2005.

[13] Lackey 2010, 11.

[14] Kelly 2008.

[15] Williamson 2002, 4.

Evidentialism is the view that 'Person S is justified in believing proposition *p* at time *t* if and only if *S*'s evidence for *p* at *t* supports believing *p*.'[16] From this falls out the claim that epistemic justification supervenes on the evidence – that there can be no change in epistemic justification without a change in the evidence. For the time being, this rough sketch is sufficient, but I shall come back to what precisely it says and does not say in section 5.

2. MOTIVATIONS FOR AN INTERPERSONAL VIEW OF TESTIMONY

This section focuses on the features that interpersonal theorists have isolated that distinguish evidential from testimonial justification, and therefore motivate a non-reductive, interpersonal view of testimony. These justify the claim that epistemic justification granted by testimony is non-evidential in nature, if by 'evidential in nature' we mean of a similar structure as the epistemic justification one gets by directly considering evidence. As a quick clarification, it is not that an act of testimony cannot be, or be presented as, evidence, but for it to be or do so would be something other than to testify.[17] The unique features of testimony qua testimony distinguish it from evidence on occasions where it is not being treated as evidence, and yet provides epistemic justification. In this there is an analogy with promising and apologising. One might present their promise as evidence that the promised action will be completed, but to do so would not truly be to promise.

2.1 INTRINSIC FEATURES OF TESTIMONIAL WARRANT

The special role for intention

Earlier, we defined testimony in terms of tellings, where to tell crucially involves intention on the part of the speaker to convey information. This is the first way in which testimonial justification should be differentiated from the justification got by considering evidence. Intention is crucial in what it is to testify; one cannot perform an act of testimony unintentionally. By contrast, intention generally undermines the justificatory powers of evidence. We tend to think of evidence which is placed so as to intentionally bring about a particular belief in a subject as 'less good' than evidence which arises naturally, even if the belief it is intended to bring about is in fact true. For example, a handkerchief dropped by accident by a killer at the scene of a crime is better evidence for that person being a killer than a handkerchief deliberately placed there to bring about such a belief, even if that person is in fact the killer. Learning that your teacher intends to induce a belief in you that *Homo sapiens* descended from *Australopithecus* does not, by contrast and in general, undermine the basis for believing them and accepting the belief. It is hard to see, if intention plays an essential role in telling, and testimony is just another form of evidence, why we should not view it as doctored, rather than genuine, evidence. If we are to view testimony as doctored evidence, it is difficult to justify the pervasive reliance we actually have on testimony in our epistemic practice. On a view which gives a particular epistemic role to the speaker's intention which is not based on evidential relations, by contrast, there is the space to explain how recognition of the speaker's

[16] Adler 2017.

[17] Moran 2005, 23.

intention can play a positive role, rather than being epistemically irrelevant, or undermining the evidential status of the belief in question.[18]

The role of responsibility

Moreover, in developing a theory of testimony, we need to account for the fact that when we tell someone something, we take responsibility in some sense for the truth of that statement. This is evidenced by the fact that blame is an appropriate response to being told a falsehood, but not towards, for example, a faulty thermometer. It is hard to ground this distinction on an evidential view where the speaker acts, much like the thermometer, as a *truth gauge*:[19] as merely an indicator of where truth could be found. The difference seems simply that in the case of a liar, it is a *person* by whom you have been deceived. This dissimilarity in the appropriate responses to being told a falsehood indicates that testimonial justification is not the same as the justification garnered by considering evidence. Though, as some critics have noted, this is not necessarily an *epistemic* difference, it is a difference nonetheless.

2.2 The structure of testimonial warrant

Non-neutral epistemic import for all parties

Regarding structure, there is a further difference between the justification granted by testimony and the justification granted by considering evidence. The epistemic justification given to all parties who consider the same evidence is the same. By contrast, the testifier and the person testified to have different epistemic relationships to the act of testimony. To demonstrate the symmetry of evidential import, consider a case where a photographer takes a picture of an incriminating scene. The photographer's relation to the photograph can be evidential, they too can learn things from its inspection. In this sense, their epistemic relation to the photograph is the same as that of the friend whom they later show it to: '[t]hey can both learn from it, or doubt what it shows.'[20] There is in this way a symmetric relation between the person who produces the evidence and the evidence and the person who receives the evidence and the evidence. They can both learn something new from consulting it, and it can serve as an independent correction of either of their initial impressions and beliefs. Contrast this with the incoherence of examining one's own testimony to establish whether or not it gives you a good reason for belief. Surely, such a chain of reasoning is possible. One might argue that they told a friend that it would rain later, people usually say things which are true, therefore it is probably true that it will rain later, but there seems something strange about it.[21] If testimony were evidence, and if the justification granted by testimony were the same as the justification granted by considering evidence, we should not have this asymmetry.

---

[18] Ibid, 18.

[19] Hinchman 2005, 563-4.

[20] Moran 2005, 10-11.

[21] There are cases where such a chain of reasoning seems more plausible, but arguably these are cases of introspection not testimony.

Revocability of the epistemic import

Lastly, there is a distinction between evidential and testimonial justification in whether or not the epistemic justification granted by the method can be revoked. When you have a piece of evidence, its evidential status is independent of any person's beliefs and intentions.[22] The person who produces the evidence cannot cancel its evidential import, and cannot alter or reduce its epistemic effect by refusing to support it. A photograph would be evidence of the existence of bigfoot whether or not the person who produces it believes it to be so. It is, then, 'in the nature of genuinely evidential relations that they are not subject to anyone's conferral or revocation.'[23] By contrast, in the case of testimony, the speaker has the exclusive authority to 'take back' or 'revoke' the epistemic status of something that they have said. For example, if I accused you of telling me that it would snow today, you would be within your rights to contend that all you testified to was that the weather would be cold. The speaker has no such authority to cancel evidential import, so it must be some other kind of epistemic influence that they have sway over.[24]

In Moran's words, '[t]aking the utterance as evidence detaches the reason giving significance of the utterance from the speaker's authority to determine what he is thereby committing himself to'.[25] A given utterance by a speaker is evidence for all sorts of things, but the speaker sees themselves as assuming a specific responsibility, and conferring a specific entitlement. You might infer from my saying that the weather tomorrow will be cold that some person has at some time uttered the word 'cold', but I am not responsible for this belief in the same way as I am responsible for the belief you have, if you accept my testimony, that the weather tomorrow will be cold. In this there is, then, an asymmetry in the relations of utterer and recipient to the instance of testimony, which is not present in cases of evidence: in the case of evidence, neither party can alter the evidential import, in the case of testimony, only the utterer can determine what it is that they assert.

3. The Dilemma

The preceding section argues that, if we are to accept that testimony has epistemic import, that epistemic import must not be evidential. There is, then, a leftover question: if the epistemic justification provided by testimony is not evidential, how is it epistemic at all? Indeed, the argument which is classically thought to devastate IVT argues precisely that the interpersonal views, if they are genuinely interpersonal, and not evidential views in disguise, deprive testimony of its epistemic import. This is Lackey's Dilemma, presented in her paper 'Testimony: Acquiring Knowledge from Others' (2011).

We might present Lackey's dilemma semi-formally as follows:

1	Testimony is both genuinely interpersonal and epistemically potent.
2	If testimony is genuinely interpersonal, then it is not epistemically potent.

[22] Ibid.
[23] Ibid, 23.
[24] Ibid, 21.
[25] Ibid, 26.

3	If testimony is epistemically potent then it is not genuinely interpersonal.	
4	Testimony is epistemically potent.	∧E 1
5	Testimony is genuinely interpersonal.	∧E 1
6	Testimony is not epistemically potent.	MP 2,5
7	⊥	¬E 4, 6

Lackey argues that the first premise should be accepted by anyone who seriously presents IVT as an account of the epistemic power of testimony. The second and third premises are supported by thought experiments. I shall briefly set out each. The crux of the dilemma is that the interpersonal theorist cannot have their cake and eat it. Either testimony is genuinely interpersonal, or it is epistemically potent, but not both. Each quality precludes the other.

First, let us deal with premise (2): if testimony is genuinely interpersonal, then it is not epistemically potent. In justifying this premise, Lackey considers Beatrice, a 'reliably unreliable testifier'.[26] Beatrice 'consistently offers assertions to her hearers that she sincerely believes to be true, but which are wholly disconnected from the truth'.[27] Imagine that Beatrice tells John that it is raining outside. Lackey argues that, because Beatrice's testimony is not connected to the truth, it is implausible to hold, as the interpersonal theorist must, that John gains epistemic warrant for his belief that it is raining outside on the basis of Beatrice's testimony. Despite Beatrice seemingly being a perfectly good candidate to give testimony on IVT, and being able to fill all the interpersonal criteria one might expect, she is unable to grant epistemic warrant through testimony. This supposedly demonstrates that interpersonal factors are not sufficient to generate epistemic justification.

Second, if testimony is epistemically potent, then it is not genuinely interpersonal. In justifying this premise, Lackey constructs an overhearing case. Her example is as follows. Abraham and Belinda think they are alone. Abraham tells Belinda that their boss is having an affair with Iris. Edgar, without their knowledge, has been listening in on their conversation. He, like Belinda, comes to believe on the basis of Abraham's testimony, which is 'both true and epistemically impeccable',[28] that his boss is having an affair with Iris.

In this case, Belinda and Edgar are stipulated to have not only the same background information, but are also 'properly functioning recipients of testimony who possess no relevant undefeated defeaters'.[29] Lackey argues that the interpersonal theorist must say that Belinda's belief in this case possesses epistemic value that Edgar's does not. She argues to the contrary that the only differences between Belinda and Edgar are interpersonal, and the only factors that might have a bearing on this difference are non-interpersonal, for example the reliability of the informant, or the amount of evidence had.

The best way to understand Lackey's dilemma seems to be as putting pressure on the idea that interpersonal features can in and of themselves be relevant to epistemic justification. In this perhaps we should understand her as asking them to give an account of what the epistemic, yet non-evidential, justification granted by testimony is. In the next section I shall give an outline of how the interpersonal theorist might respond to this challenge.

[26] Lackey 2010, 13.
[27] Ibid.
[28] Ibid, 15.
[29] Ibid.

PART 2: HOW IS NON-EVIDENTIAL TESTIMONIAL WARRANT EPISTEMIC?

The interpersonal theorist's response to Lackey's challenge comes in two steps. First, they must clarify exactly by what mechanism testimony justifies beliefs. In this, I am going to advocate taking a 'buck-passing' approach to testimony, developing work done by Judith Baker and Philip Clark (2018). Then, the interpersonal theorist must establish how exactly that mechanism, buck-passing, might grant *epistemic* justification. I argue that there are already-existing distinctions in the field of epistemic justification which they can use to explain this.

4. STRONG BUCK-PASSING AND IVT

First, by what mechanism does testimony grant justification to our beliefs? If testimony is to be a *sui generis* form of justification, then it must be associated with a *sui generis* mechanism of generating said justification. If it were to grant justification in the same *way* as traditionally evidential sources of justification, it would seem likely that it should grant the same *kind* of justification, namely evidential. There must be, then, some characteristically interpersonal principle of justification. One suggestion has been that the interpersonal theorist should endorse an epistemic principle of buck-passing. This has been recommended by Baker and Clark, who use a notion called 'strong B-P':

> When challenged to produce the evidence that justifies her belief that *p*, *A* can acknowledge that she is unable to do so by herself, without help from her source, without thereby undermining her claim to know that *p*.[30]

My aim in this section is to precisify the account given by Baker and Clark, to raise and dismiss some additional concerns, and to show how to incorporate the traditional interpersonal accounts on such a picture.

If strong B-P is sometimes true, then although *A* does not herself have sufficient evidence, she still has knowledge: she can defer the challenge for evidence to her testimonial source. It is, therefore, sometimes the case that an agent can have epistemic justification without having direct evidential justification, and the interpersonal theorist renews their claim that epistemic justification needn't be directly evidential. As an example, consider the following. Bernie tells Alice that it is raining. This is the only justification that Alice has that it is raining, and in normal circumstances, when there is no reason to suppose any kind of trickery in play, we would take this to be sufficient for Alice to know that it is raining. (This is putting to one side, as is done for the whole essay, any kind of sceptical concern.) Charlie asks Alice how she knows that it is raining. She can defer the epistemic challenge to Bernie, and say that she knows because Bernie told her, and that if Charlie wants further justification she should ask Bernie, despite not having any evidence, and without thereby giving up her claim to know.

The example helps demonstrate that the justification granted by Bernie's testimony for Alice's belief that it is raining has the same features which distinguish it from the justification got by considering evidence as those earlier isolated. There is a particular role for intention: for Bernie's utterance to grant epistemic justification to Alice's belief, he must intend it to do so, and this intention strengthens rather than weakens the epistemic value of the utterance. There

[30] Baker and Clark 2018, 185.

is a particular role for responsibility, in that Bernie clearly takes responsibility for the truth of Alice's belief, if she is able to defer challenges back to him, and it would be appropriate to take the reactive attitude of blame towards him were the belief in fact to be false. In fact, it seems to be the very fact of Bernie's taking this responsibility that grants Alice's belief the epistemic justification that it enjoys. Further, it wouldn't make sense to say that Bernie could gain epistemic justification from the fact of his own testimony, or that he could gain epistemic justification from the fact of his testimony in the way that Alice does.

Adopting strong B-P is therefore incompatible with the view that testimony is just another species of evidence: if testimony justifies a belief by giving the recipient of the testimony an epistemic right to 'pass the buck', then this mode of justification has features which differentiate it from getting epistemic justification by considering evidence directly. The interpersonal theorist must hold that the warrant granted by testimony is different from the warrant granted by considering evidence directly at least to some extent *because* strong B-P is true only of testimonially justified beliefs. If this principle is only true of testimonially justified beliefs, if it is only in the case of testimony that you have an epistemic right to defer challenges to the epistemic source without giving up your claim to know, then the justification that testimony grants beliefs is distinct from the justification granted by considering evidence for this very reason.

First, let us justify the claim that the right to defer associated with strong B-P is unique to testimonial knowledge. There is a weaker version of the principle, which is available to someone who considers testimony as evidence, and which is not testimony-specific. This is a principle that Baker and Clark call 'weak B-P'; it is as follows:

> Having cited her source and given her reasons for thinking the source is reliable, A can tell her interlocutor (and thereby express a justified belief) that if he wants more evidence that *p* he should go ask the source, without thereby undermining her claim to know that *p*.[31]

There is some sense in which the person who views testimony as evidence can acknowledge, in certain situations, an epistemic right to pass the buck. Weak B-P is applicable generally to all epistemic sources. For example, this buck-passing principle can be applied when the source of epistemic warrant for A is looking at a photograph. A could tell her interlocutor that she believes that *p* on the basis of seeing a photograph that appears to show *p*, and she believes that the photograph has not been tampered with, and that if the interlocutor wants more evidence he should consult the photograph himself.

However, there is a key difference between this weaker principle, and the more powerful Strong B-P. In the weak principle, this 'move of last resort' is 'epistemically appropriate only given that the recipient of the testimony … has *already* justified her reliance on that testimony'.[32] It is only because, for the reductionist, A has reasons to think the photograph a reliable source of information that she can 'pass the buck' back to the photograph. The reductionist can therefore say that A is justified in believing the testimony *because* A has reasons to believe that the assertion is a 'reliable indication of the truth of *p*'.[33] By contrast, it is only in the case of Strong B-P that the recipient of the testimony can acknowledge that they

[31] Ibid, 182.

[32] Goldberg 2006, 135-136.

[33] Baker and Clark 2018, 183.

personally do not have access to the evidence without thereby undermining a claim to knowledge. It's hard to see how this difference might carry over to the case where A gains her epistemic justification by directly considering evidence. Outside of a testimonial exchange, it is hard to see how a subject may avoid needing to have enough evidence to justify their belief to count as having knowledge.[34] If the only way that they can epistemically justify their belief is through directly considering evidence, and they concede that they do not have the evidence to justify their belief, then the belief cannot be epistemically justified. Without endorsing the idea that testimony provides a form of epistemic justification which is distinct from the epistemic justification that is provided by directly considering evidence, then, one cannot endorse Strong B-P. Strong B-P is unique to testimonial justification.

In order to rescue their position from Lackey's dilemma, however, the interpersonal theorist needs B-P to not only be feasible and unique to testimony, but also sometimes true. It should not just be a plausible epistemic principle, but a principle which is sometimes reflected in practice. Benjamin McMyler (2011) argues that this is in fact the case, observing that a hearer has an epistemic right to defer challenges back to the original speaker. When the recipient is challenged to defend a claim accepted on testimony, and cannot immediately do so, the hearer need not immediately give up their claim to know. Instead, they can defer the challenge back to the original speaker. If the original speaker cannot meet the challenge, the hearer does indeed have to give up their claim to knowledge. This is supported by the plausibility of the example given above. If asked how you know that the creatures that exist today evolved from earlier organisms by a process of natural selection, you can defer the epistemic challenge to your biology teacher by saying that they taught you that, without thereby giving up your claim to know. The key point here is that the question of knowledge is not *settled* by the fact that the hearer cannot defend the belief on their own. In fact, we treat the hearer's inability to meet the challenge as inconclusive. Though one might disagree with McMyler on the specifics, for example, that the recipient must give up their claim to knowledge if their informant cannot meet the challenge, the observation remains that there is an epistemically legitimate practice of deference that the interpersonal theorist can make use of.

5. Locating the interpersonal theorist in epistemic territory

Having established that buck-passing is a suitable candidate for the interpersonal relationship in question, we must now address the issue of how buck-passing generates a justification which is distinctively epistemic. By contrast, one might instead think that buck-passing generates a justification in the sense that it is morally, conventionally, or socially justified to hold the belief in question, that no one would blame you for holding it, but that this kind of justification does not give you knowledge. We could construct an argument from it being obvious that testimony grants epistemic justification, and that testimony can be adequately thought of in terms of buck-passing, to, *a fortiori*, the conclusion that buck-passing can grant epistemic justification. However, this claim, if true, is so important for our theory of epistemic justification more generally, that it is worth elucidating exactly how it achieves such an effect.

There are two ways that the interpersonal theorist might argue this case to diffuse the force of Lackey's dilemma. First, they could develop a theory of epistemic justification which

[34] Ibid, 186.

is distinct from the evidentialist position, and find an explanation of how buck-passing grants epistemic justification within that framework. Second, they might argue that, in fact, there is room for an explanation of epistemic justification granted by passing the buck within the epistemic framework that the evidentialist adopts. Since this latter option is the better response to Lackey in particular, and has interesting consequences for our theories of epistemic justification more generally, I direct my attention hence.

Let us start by dividing up the epistemic territory. Here I shall draw upon work by Hamid Vahid (1998), who distinguishes deontological and nondeontological theories of epistemic justification. Deontological theories construe epistemic justification as fulfilment of epistemic duty, whatever that epistemic duty might consist in. If the agent fulfils their epistemic duty with regard to the belief, that belief is justified. Nondeontological theories, by contrast, deny this claim. The most famous variant of nondeontological theories involves truth-conducive justification. The most famous instance of this variant is likely reliabilism, proponents of which hold that a belief is justified just in case it is formed by a belief-forming process that produces 'mostly true beliefs or a high ratio of true beliefs to false beliefs.'[35] A difficulty arises when thinking about where to fit evidence and evidentialism into this picture. Either one could say that our epistemic duty is to apportion our beliefs to the evidence, or that our beliefs are justified just in case they are apportioned to the evidence, regardless of what our epistemic duty consists in. These two theories could well be extensionally equivalent: all the same beliefs might count as justified, but the beliefs would be justified for different reasons. The deontological evidential account of epistemic justification is what I suggest the buck-passing interpersonal theorist endorse in response to the concerns raised by Lackey.

By its status as a deontological account of epistemic justification, we can say that a belief is epistemically justified just in case the believer has fulfilled their epistemic duties with regard to the belief. Then, by introducing the evidential component, we say that the epistemic duty in question is a duty to apportion your beliefs to the evidence. To show how this kind of theory of epistemic justification fits with the interpersonal view, it's helpful to make an analogy with promising. It seems fair to say, for example, that one has a duty to protect their small child, but if they pass on their child to a caregiver who promises to protect the small child, they have discharged this duty. They fulfil their duty to protect the child by 'passing the buck' onto another caregiver. Similarly, the buck-passing interpersonal theorist might say, the epistemic agent can fulfil their epistemic duty to apportion their beliefs to the evidence by allowing another person to take on this responsibility. In order to have knowledge, the agent must fulfil their duty to apportion their beliefs to the evidence, but if someone else testifies to the truth of the belief, and the agent accepts this testimony, they discharge this duty to them. There is therefore more than one way in which to apportion one's belief to the evidence; it can be done either directly, by gathering evidence and generating evidential warrant, or indirectly, by taking the word of others.

On the deontological evidential variant, our epistemic duty is to apportion our beliefs to the evidence, but we do not prescribe how this apportioning should be achieved. On such an account, though we maintain that testimony is not itself evidence, we deny that this claim is inconsistent with endorsing any form of evidentialism. We can endorse both the claim that we obtain epistemic justification by apportioning beliefs to the evidence and the claim that testimony provides a *sui generis* mode of epistemic justification, if we endorse a theory of

[35] Becker 2020.

epistemic justification which has the scope to allow for different methods of apportioning one's beliefs to the evidence. There are different ways of fulfilling one's epistemic duty with regards to the belief.[36]

This way of conceiving of how testimony provides epistemic justification, as well as showing how one can have epistemic justification which is not evidential, allays some of the worries that might be raised by Lackey's counterexamples. The first was the case of Beatrice, the reliably unreliable testifier. Lackey argued that Beatrice should be a prime candidate for granting testimonial justification interpersonally, but her guarantees were completely disconnected from the truth. The interpersonal response to this counterexample should now be that, though Beatrice might be able to grant some epistemic warrant, if there is enough contrary evidence, this can outweigh the epistemic warrant granted by Beatrice. Contrast two cases. In one case, Fred, an ordinary epistemic agent, has never met Beatrice, and knows nothing of her unique epistemic tendencies. They meet, and Beatrice tells him that it's raining outside. It seems, in the absence of further evidence, Fred would be justified in his subsequent belief that it is raining outside, given that he accepts the testimony. Compare this to a case where Fred knows full well what Beatrice is like. He has seen her produce reliably unreliable testimony on many occasions before. In this case, because the evidence he has outweighs any kind of epistemic warrant that Beatrice might grant him by her testimony, he would not be justified in forming a belief on its basis. The overall obligation is to apportion one's beliefs to the evidence. Testimony is just one way, a particularly distinctive way, by which this can be achieved.

There are two other ways, briefly, that the buck-passing interpersonal theorist might respond to the Beatrice case. Firstly, they might place the weak and intuitively plausible requirement on the person giving the testimony that they be an epistemic agent. This means that they must at least be reasons-responsive to be able to make the sort of guarantee that the interpersonal theorist is dealing with. In the Beatrice case, by contrast, Beatrice will believe incorrectly regardless of the reasons present for or against. This is weaker than Hinchman's restriction on the testifier, that they be in some sense 'truth-tracking'. Instead of tracking the truth, all that is required is that relevant reasons/evidence 'make a difference' to the testifier, that they be the sort of agent that could be in the business of sorting truth from falsehood. Compare this case with trusting the assurances of a very small child, someone who is radically deluded (which is what we might think of Beatrice), or of an animal, e.g. your pet dog pandering at you to feed it. In none of these cases would we think an epistemic warrant conferred, because the person, or animal, offering the assurance is not an epistemic agent: they are not reasons-responsive, in the sense of being in the business of sorting truths from falsehood. This is analogous with a similar restriction on promising. Something is only capable of making a promise if it is a moral agent; something is only capable of making an epistemic guarantee if it is an epistemic agent.

Second, they might argue that, though testimony can transmit knowledge, it cannot generate it. This means that at the beginning of a chain of testimony there must be someone who independently knows the fact in question. This would rule out the Beatrice case, as she is

[36] To reiterate, this is one of two routes the interpersonal theorist might take. They might alternatively decide that evidentialism holds no water in any respect, and give a competing account of what our epistemic duty might consist in. Though this alternative is available, the endorsed account is sufficient to diffuse the force of Lackey's dilemma, and there are at least some benefits to connecting testimony to evidence in some respect, though the interpersonal theorist maintains that the two are not the same.

the beginning of the testimonial chain and, radically deluded as she is, she intuitively cannot be said to have knowledge.

Lackey's second counterexample was the overhearing case. The interpersonal theorist now also has a rejoinder to this. They state that, though Belinda has some warrant that Edgar doesn't have, this doesn't rule out, what is explicitly specified in the case as it is given, that both plausibly have enough evidential warrant for the belief to be independently justified. Both are justified in their belief that Iris is having an affair with their boss, because Abraham is specified to be a good source of evidence, but Belinda is better justified than Edgar is, because if it wasn't known to both of them that Abraham was reliable, then Belinda could remain justified in her belief when challenged by deferring the challenge back to Abraham, but Edgar couldn't. Edgar can use the fact that Abraham told Belinda that Iris is having an affair with their boss as evidence, but he cannot pass the buck back to Abraham if challenged to justify his belief. His belief that *p* is justified in virtue of the fact that Abraham has told Belinda that *p* and Abraham is known to be a reliable source of evidence, but if this were not sufficient evidence, he would not know that *p*. If Abraham's speech act is evidence, then in virtue of his relation to that evidence directly, and his knowledge that Abraham is reliable, Edgar can justify a belief which Abraham asserts to be the case. However, Edgar is only related to that evidence. If he did *not* know that Abraham was reliable, he might not know that *p*. By contrast, in virtue of accepting Abraham's adoption of epistemic responsibility for his belief, Belinda becomes indirectly related to all the evidence that Abraham has, but which he has not shared. Even if Belinda were not directly related to enough evidence to justify a belief that *p*, Belinda might know that *p*, because she is indirectly related to the evidence that Abraham has, in a particular way which fulfils her epistemic obligations. In her case, but not Edgar's, has Abraham taken responsibility for his belief's truth; in her case, but not Edgar's, has the relevant epistemic obligation been fulfilled.

Each of these responses is consistent with the claim that the kind of epistemic warrant granted by testimony is not the same as the kind of epistemic warrant granted by considering evidence directly, and also consistent with the claim that an epistemically justified belief is one that is justified by fulfilling your epistemic duty to apportion your beliefs to the evidence. There is no clear conflict between the interpersonal view of testimony and what is, at root, a very evidentialist way of viewing how beliefs are epistemically justified. Testimony is not evidence, in contrast to other ways of being indirectly related to evidence which are evidence, for example photographs, and yet relates you to evidence in such a way that being so related allows you to fulfil your epistemic responsibility with regard to the belief. This particular way of being related to the evidence is distinctively social, it arises only as a result of relationships between epistemic persons, and it is the relationship itself which bears the epistemic weight.

So, how does this view square with the central thesis of evidentialism we considered earlier? Evidentialism, as defined above, is the claim that 'Person S is justified in believing proposition *p* at time *t* if and only if *S*'s evidence for *p* at *t* supports believing *p.*'[37] This is consistent with how the interpersonal view is here interpreted only so long as we include within S's evidence the evidence that they are indirectly related to on the basis of another person's testimony. Though the person's speech act on its own might not be enough to epistemically justify belief, by accepting the conferral of epistemic responsibility, by accepting a right to 'pass-the-buck', the agent becomes indirectly related to evidence which the testifier has, but

[37] Adler, 2017.

which the agent does not. The agent is then justified in their belief, provided that the testifier is justified in their belief, without having to ascertain anything more about the testifier's reliability, or their credentials as a source of evidence. This, however, is a slightly strange way to interpret what S's evidence is. The interpersonal theorist might instead see reason to depart from evidentialism altogether.

PART 3: CONCLUSION AND CONSEQUENCES

Though we can interpret IVT in such a way that it is consistent with evidentialism, doing so has significant consequences for our theory of epistemic justification more generally. In the most interesting case, we are left with the claim that the relationships developed between epistemic agents can themselves bear epistemic weight, in a way which does not merely reduce to their status as evidence for a particular belief. In short, we make space for something which looks very much like genuinely social, collective knowledge. This allows for the epistemic agent to be justified in a belief, even without themselves having evidential reasons for the belief's likely being true. We deny that 'all reasons to believe *p* are evidence for *p*'[38]: that a given person testified to *p*, and you accepted their testimony, gives you a reason to believe *p* which does not reduce to the testimony itself being evidence for this belief. The interpersonal theorist here described therefore holds that, though a belief is epistemically justified in virtue of an epistemic duty being fulfilled, where that epistemic duty is one of apportioning beliefs to evidence, the 'reasons that underwrite an audience's believing a speaker, and in that way believing what they are told, cannot be of an evidential kind'.[39]

What IVT shows is that epistemic justification is best understood in terms of fulfilling our epistemic duties, where the epistemic duty in question is to apportion beliefs to the evidence. This is in contrast to an understanding of epistemic justification in terms of apportioning beliefs to the evidence directly. This allows for a greater scope of ways in which beliefs can be epistemically justified. One of the desirable consequences of this picture is that on it we may allow for social relations to genuinely bear epistemic weight, in a way which does not reduce just using other people as sources of evidence. By allowing epistemic justification to be generated by interpersonal relations, we account for genuinely social, collective knowledge.

Therefore, if the interpersonal theorist has correctly described some key differences between testimonial justification and evidential justification, and testimony is indeed a source of epistemic justification, then we must rethink some aspects of our theory of epistemic justification generally. Lackey's dilemma is unsuccessful, because we can pair the interpersonal view with an evidentialist view if necessary, but even if we maintain a broadly evidentialist framework, we must rethink the ways in which we can be related to evidence to garner epistemic justification. This allows us to account for the genuinely social features of testimony, and to make room for collective, genuinely shared, interpersonal knowledge. On the particular view described, which merges the evidentialist and interpersonal views, we get the best of both worlds. Testimony is still connected to the evidence, but there is space for a *sui*

[38] Way 2016, 805.

[39] Longworth 2020, 271.

generis species of testimonial warrant. I can be justified in a particular belief on the basis of evidence that another person has.[40]

REFERENCES

Adler, J. (2017). Epistemological Problems of Testimony. *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.). E.N. Zalta (ed.)

Baker, J. and Clark, P. (2018). Epistemic buck-passing and the interpersonal view of testimony. *Canadian Journal of Philosophy*, 48(2): 178-199.

Becker, K. (2020). Reliabilism. *The Internet Encyclopedia of Philosophy*.

Green C.R. (2020). Epistemology of Testimony. *The Internet Encyclopedia of Philosophy*.

Hinchman, E.S. (2005). Telling as Inviting to Trust. *Philosophy and Phenomenological Research*, 70: 562-587.

Kelly, T. (2008). Evidence: Fundamental Concepts and the Phenomenal Conception. *Philosophy Compass*, 3: 933-955.

Lackey, J. (2010). Testimony: acquiring knowledge from others. In Alvin I. Goldman & Dennis Whitcomb (eds.), *Social Epistemology: Essential Readings*. Oxford University Press.

Longworth, G. (2020). Corresponding reasons: on Richard Moran's The Exchange of Words. *Tandf: Philosophical Explorations* 23 (3):271-280.

McMyler, B. (2011). Knowing at Second Hand. In *Testimony, Trust, and Authority*. Oxford University Press.

Moran, R. (2005). Getting told and being believed. *Philosopher's Imprint* 5(5): 1-29.

Vahid, H. (1998). Deontic vs. Nondeontic Conceptions of Epistemic Justification. *Erkenntnis* 49: 285-301.

Way, J. (2016). Two Arguments for Evidentialism. *The Philosophical Quarterly* 66(265): 805-818.

Williamson, T. (2002). Evidence. In *Knowledge and its Limits*, by Williamson, T. (Oxford, 2002; pubd online Nov. 2003). Oxford Scholarship Online.

[40] I would like to thank the Editor of *Critique*, as well as two anonymous reviewers, for their feedback which has greatly aided the readability of this paper. I would also like to thank Jessie Munton for her help throughout the writing process.

What Are Hegel's Metaphilosophical Views?

CHEONG KWANG AIK ELDRICK
National University of Singapore

INTRODUCTION

Hegel's metaphilosophical views – his philosophy of philosophy – are not as clearly set out as his other views on consciousness, the mind, social and political philosophy, in the sense that he did not dedicate a single work entirely elucidating his conception of philosophy.[1] However, I would stipulate that his metaphilosophical views play a crucial role in understanding his philosophy itself, as they form the basis of his motivation for treating, for instance, human consciousness in the manner that he does i.e. as spirit. In this essay, I elaborate on Hegel's arguments in the 'Preface' and 'Introduction' to the *Phenomenology of Spirit* (*PhG* from hereon) and the 'Preface' and 'Introduction' to the *Science of Logic* (*Logic* from hereon). I address three main topics: (1) Philosophy as the domain of thought; (2) Philosophy as the embrace of contradictions; and (3) Philosophy as the discipline to resolve these contradictions. For this essay, I will treat point 1 and 2 in detail, while treating the last point merely as a corollary that follows from the first two points.

1. PHILOSOPHY AS THE DOMAIN OF THOUGHT

Hegel's thoughts on human thought are part of his effort to carve out a 'scientific' system that explains how we come to perceive and conceive of reality around us. Therefore, before I talk about how Hegel thinks of philosophy as the domain of thought and why he might have believed that understanding how we perceive or understanding 'what knowing is'[2] is important in itself for philosophy, I shall clarify Hegel's conception of science.

1.1 HEGEL'S CONCEPTION OF SCIENCE (*WISSENSCHAFT*)

Hegel's *Wissenschaft* refers to a systematic study of things (or subject matters) rather than solely the *a posteriori* study of natural facts or phenomenon. As such, the study of anything can be a science as long as it is studied from a systematic point of view, with close attention to system-building. The notion of systematicity is broad but for this essay, I shall characterise it as the intellectual practice of conceptualising and demarcating things under one grand vision.[3]

[1] However, Hegel did write a few articles addressing philosophy as a discipline with relation to his time. See Giovanni and Harris 1985
[2] McDowell 2018
[3] It is probably for this reason that the 20th century philosopher Isaiah Berlin categorised Hegel as a "hedgehog" who "relate everything to a single central vision, one system, more or less coherent or

Hegel believed that logic was one such system to be had and therefore, one such science, which partly explains why he named his book on the subject the *Logic*. The other reason stems from his views on the state of logic during his time. In the *Introduction* of the *Logic*, Hegel laments that 'the need for a reformation of logic has long been felt. In the form and content in which it is found in the textbooks, it must be said that it has fallen into disrepute.'[4]

Hegel was referring to the traditional formal or general logic that was then dominant, with its emphasis on deduction (or syllogism), although induction was also part of his system.[5] Hegel does not decry the entirety of this (otherwise commonly-known-as) Aristotelian logic. Rather, he can still be seen as pursuing the strands of Aristotelian logic but in a different direction, in what can be called an ontological rather than 'semantic [or] syntactic' manner.[6] It is in this manner that Hegel takes it upon himself to revive the scientific nature of logic, as we shall see in more detail below.

In his exploration of human cognition and the various stages it undergoes in the *PhG*, the dialectical movement that spirit 'experiences' prior to reaching Absolute Knowing and the subsequent development of this movement in *Logic* reveals this scientific character. However, in *Logic,* the same dialectical movement manifests, although in this case, the notion of the subject-object relation has been overcome. The difference between these concepts is that spirit is Hegel's logic in action. It is a concrete manifestation of the logic, but with particular focus on human cognition and understanding of the world.[7] Seen in this light, a characteristic scientific hallmark is retained throughout Hegel's system, whether with respect to consciousness as represented by spirit or Hegel's conception of logic.

In short, labelling Hegel as unscientific from the vantage point of science in the more modern sense of the term i.e. the *a posteriori* study of natural facts or phenomenon would be a cartographic mistake. For Hegel, keeping to 'science' in the sense of building a system to account for phenomena is very important.[8] Therefore, his system of analysing human cognition and understanding as well as the realm of pure thought, as shown in the *Logic*, represents the peak of scientific exercise during his time.

1.2 System Building – Hegel's Dialectic

Having clarified Hegel's conception of science, I shall proceed to expound on Hegel's treatment of Philosophy as the domain of thought. In particular, I shall look at how Hegel treated human cognition in a scientific manner as stated above i.e. with a pressing emphasis on system building. The main questions addressed in this section are: (a) What is Hegel's dialectic? and (b) What does his dialectic say about his conception of Philosophy?

---

articulate, in terms of which [he] understand[s], think and feel – a single, universal, organising principle in terms of which alone all that they are and say has significance…" Berlin 2013, 2

[4] Hegel 2010

[5] Smith 2020

[6] Redding 2007

[7] I use "cognition" and "understanding" in their ordinary sense, not in the specific sense that Hegel attaches to them or their Germanic root attaches to them. For the latter, see Hegel 2018, 326, 327

[8] See Hegel 2018, 36-7 S76-7 where Hegel differentiates between different kinds of knowledge. In some respect, Hegel's view of 'science' belongs to the older, scholastic, tradition. His philosophy of nature, for example, re-works Aristotle's Physics in which the guiding concept is of substance unfolding itself with teleological necessity. (From correspondence with my supervisor, Prof Luke O' Sullivan)

(A) WHAT IS HEGEL'S DIALECTIC?

The *PhG* is an attempt by Hegel to explore every aspect of human consciousness as it progresses via a hierarchical and logically sequential order. Human consciousness, in the form of spirit, proceeds from the most basic form of "Sensory Certainty" to "Perception" to what he calls "Absolute Knowledge". Within this transformation, Spirit undergoes a sequence that is commonly labelled the dialectic or what Hegel would call "the presentation of the course of experience".[9] As Charles Taylor puts it, "The course of Geist's [Spirit] development towards self-knowledge lies through the initial confusions, misconceptions and truncated visions of men." (Taylor 1977, 127)

Spirit first assumes one of the "shapes of consciousness" (Hegel 2018, 18 S36) of Sensory Certainty, for instance, but then recognises contradictions or oppositions that naturally arose. Given this outcome, Spirit overcomes itself and takes on another shape — Perception. It undergoes the very same process of recognising its own limitation in the shape it has taken and overcomes itself towards the next shape in the hierarchy. Therefore, the make-up of Hegel's dialectic in the *Phenomenology* is this hierarchical and logically sequential order which in turn characterises Spirit's own movement. However, if this movement is what characterises Spirit's main motive towards change, what is it that causes it to proceed from one elevation to another? How does Spirit recognise the contradiction that arose? Hegel discusses this briefly in the *Introduction* of the *Logic*: 'What propels the concept onward is the already mentioned negative which it possesses in itself; it is this that constitutes the truly dialectical factor.' (Hegel 2010, 34)

The contradiction that causes Spirit to recognise its limitations is inherent in all the "forms of consciousness" prior to Absolute Knowledge. They lie dormant and are realised only when Spirit fails to achieve the "purpose it is bent on realizing or [a] standard it must meet." (Taylor 1977, 131) Subsequently, in the process of sublation (*Aufhebung*), the contradiction is not negated and done away with. Instead, the contradiction is embraced and manifested as part of the accumulating whole of Spirit. This embrace of contradiction is a necessary process that cannot be absent from the dialectic as a whole. (It is paramount to Hegel's conception of Philosophy and would be explored in greater details in Section 2)

In short, Hegel's dialectic[10] is a dynamic process of thought, organising itself in a hierarchical and logically sequential order. At the same time, its movement is spurred by an innate contradiction that lies within each form of consciousness that causes it to move to the next stage in the hierarchy. In the midst of this elevation, contradiction is embraced, not negated or destroyed. Therefore, Hegel characterises, via a systematic route, how human consciousness proceeds and therefore, conducts a thorough philosophical investigation in the domain of human thought.[11] In doing so, he carries out his task of understanding human cognition itself before tackling the first-order questions within philosophy.

---

[9] I treat this as being what phenomenology essentially is in that the focus is on understanding or revealing the depths of experience beyond how it appears to us. See Hegel 2018, 41, S87

[10] See Kojeve 1980, 183-4 for a different reading where Kojeve talks about how Hegel "knowingly abandon Dialectic conceived as a philosophical method."

[11] I am not presupposing that Hegel's arguments are valid. Rather, what I am doing here is outlining Hegel's agenda as a philosopher concerned with philosophy as a domain of thought. For criticisms, see Pippin 1993

(B) WHAT DOES HEGEL'S DIALECTICAL FRAMEWORK SAY ABOUT HIS CONCEPTION OF PHILOSOPHY?

Given Hegel's emphasis on analysing how we perceive and conceive the world, surely it is a tautological truth that philosophy is the domain of thought. As Graham Priest puts it, 'no one before this century tried harder than Hegel to think through the consequences of thought thinking about itself, or of categories applying to themselves.'[12] Given this, how then would Hegel think one should approach philosophy? In other words, what does Hegel's dialectical framework say about his conception of philosophy?

Firstly, we should note that Hegel holds thought to be a necessary and sufficient capability of human individuals in understanding the world i.e., there is nothing above and beyond human experience in grasping the world. However, this does not mean that anyone (on the basis that everyone thinks) can do philosophy. The culmination of the dialectic in Absolute Knowledge suggests that to engage in philosophy proper, we must be situated in a right frame of mind. By this, I am not referring to a sound mental health or a mind free from neurosis. Instead, I am referring to Hegel's spirit after it has taken on the "shapes of consciousness" of Absolute Knowledge.[13]

Prior to this undertaking, philosophy itself must encompass the task of looking at human cognition as the latter ponders philosophical problems, as Hegel states in this passage: 'It is a natural idea that in philosophy, before we come to deal with the Thing itself, namely with the actual cognition of what in truth is, it is necessary first to come to an understanding about cognition, which is regarded as the instrument by which we take possession of the absolute, or as the medium through which we catch sight of it.' (Hegel 2018, 35, S73)

As concepts arise from our thoughts, a flawed way of looking at things would result in a flawed concept.[14] This is apparent from Hegel's objection against Schelling in paragraph 16 of the *Preface* of the *PhG*: 'this formalism presents this monotony and abstract universality as the absolute; it assures us that dissatisfaction with it is an incapacity to master the absolute standpoint and stick to it.' (Hegel 2018, 10) Hegel accuses Schelling's concept of the Absolute of being a "formalism" that does not allow differences, thereby subsuming everything as one and the same. Seen in the light of Hegel's metaphilosophical views being that human cognition is the basis of which philosophical insights emerge, this indicates that Hegel thought Schelling had a flawed conception of the Absolute because it had been derived it from a flawed kind of thinking. The way to correct this flaw is then to look at what constitutes "good" thinking or the aforementioned right frame of mind.

At this point, I would like to address an objection by the philosopher John McDowell to Hegel's agenda as it is directly related to the point I am making about Hegel's metaphilosophical views. In *What is the Phenomenology About?* McDowell comments that Hegel, rather than trying to work out what exactly is human cognition prior to actual

[12] Priest 1989. This quote can be read in multiple ways, but I would generalise over the details and treat it as if Priest is talking about how Hegel pays careful attention to thought, rather than about thought thinking about itself in the metaphysical sense.

[13] See Hegel 2018, 41-2 – 'This is why the moments of the whole are shapes of consciousness. In pressing on to its true existence, consciousness will reach a point at which it sheds its semblance of being burdened with alien material that is only for it and as an other, a point where the appearance becomes equal to the essence, where consequently its presentation coincides with just this point in the authentic science of spirit; and finally, when consciousness itself grasps this its essence, it will signify the nature of absolute knowledge itself.'

[14] One such flawed thinking that Hegel eschews is common sense. See Forster 2019, 52

philosophy, is in fact eschewing this task and getting straight to the science of consciousness: 'I want to stress here…that what Hegel is doing in these opening paragraphs [i.e. passage 73] of the Introduction is rejecting the natural supposition that philosophy should address the question what cognition is before engaging in cognition.'[15]

McDowell puts forth argument in support of his claim that Hegel was trying to work out a "Science of the Experience of Consciousness", rather than analysing human cognition as a pre-philosophical or metaphilosophical exercise. As much as I agree with McDowell that Hegel was trying to work out a science of human consciousness as fast as he could, I believe he also found it necessary to look at human cognition itself, as it is only though this endeavour that we can get hold of the 'cloud of errors'. [16]Science itself cannot 'come on the scene' without our having understood how we can even conceive of science as system-building in the first place, which is to say without our having engaged in a prior study of what knowing is. Furthermore, if we do not undertake this task, science as it unfolds could be flawed due to our thoughts being flawed. Therefore, in order to ensure that the philosophical problems that come after this initial study are genuine problems rather than pseudo-ones, I maintain that Hegel did indeed regard a normative metaphilosophical preoccupation as a requirement of a scientific system of thought.[17]

To conclude, Hegel holds philosophy to be the domain of human thought. His dialectic is put into practice, as per any scientific trials or experiments, in understanding the phenomenon of human consciousness as seen in the *PhG*. At the same time, given that philosophy is of this nature, Hegel holds it to be fundamentally important for us to understand the nature of our subjective conception of the world in order to ensure that philosophical problems posed are genuine problems rather than nonsense. To make philosophy a science while not neglecting its inclinations towards the armchair i.e. predominantly *a priori* thinking is a tall order that besets Hegel. In order to elevate philosophy from 'just knowing' to 'actual knowing'[18], analysing human cognition is necessary and this endeavour itself is part of Hegel's metaphilosophical endeavours.

2. Philosophy as the embrace of Contradiction

Hegel embraces contradiction and thinks that Philosophy should do too. Hegel's views on the notion of contradiction are primarily exemplified in his questioning of the sacrosanct law in Classical Logic, namely the Law of Non-Contradiction (LNC) - The proposition '"It is true and not true" is always false.' Whether or not Hegel actually rejects LNC is a contentious

[15] McDowell 2018
[16] Hegel 2018, 35, S73
[17] Once again, the first paragraph (§73) of the Introduction says a lot about Hegel's views on human cognition and our grasping of truth and knowledge. If my interpretation is correct, Hegel thinks that it is crucial to look at human cognition since if philosophy is the domain of human thought, only the right type of cognition would ensure the truth. ("For, if cognition is the instrument for gaining possession of the absolute essence, it is immediately obvious that the application of an instrument to a Thing does not in fact leave it as it is for itself, but rather effects a forming and alteration of it.") On the other hand, even if human cognition were just an unchangeable medium via which the truth is fed to us, a realization of this fact can help us to evaluate the degree of legitimacy of those truths. ("Or if cognition is not an instrument of our activity but a sort of passive medium through which the light of truth reaches us, then again we do not receive the truth as it is in itself, but only as it is through and in this medium.")
[18] Hegel 2018, 5, S5

claim, and I would proceed to discuss what other scholars have said about it, as well as to incorporate my own views on the matter.

LNC existed throughout the history of Logic from Aristotelian Logic to the modern Fregean and Russellian logic (or what is now known as predicate or first-order logic) and its truth is taken to be absolutely necessary. That is to say, if we were to model truths as possible world, LNC would hold in every possible world. It is for this reason that the early twentieth century so-called Analytic Philosophers such as Russell and Moore rejected Hegel. They wondered how it was possible to have a 'true contradiction'.[19] The antagonism towards Hegel lies in his embrace of contradiction, as opposed to a mechanical and binary truth-functional approach towards Philosophy. As Hegel made known in passage 15 of the Preface to the *PhG*:

> When the knowing subject parades this single immobile form around in whatever is at hand, when the material is dipped into this stagnant element from outside, this does not fulfil what is needed any better than arbitrary notions about the content. It is not, that is, the wealth of shapes surging up from itself and their self-determining differentiation.[20]

The argument against Fichte's 'monochromatic formalism' is due to Hegel's rejection of the consequence Fichte drew from his absolute principle (which is deemed "the unity of self-consciousness").[21] Fichte's claim that whatever is affected by contradiction destroys the unity of self-consciousness and hence, the law of non-contradiction is the 'highest standard of thought and reality'[22] The contrast between the kind of monotonous thinking of Fichte and Hegel's own views can be seen from Hegel's rendering of the former's thoughts as "immobile", "stagnant" as compared to in Hegel's view, the needed or required "wealth of shapes surging up from itself and their self-determining differentiation". The emphasis on movement or flux, that stems from the notion of differentiation i.e. something is itself only when it can be separated from something else is taken by Hegel to be of paramount importance. As such, contradiction is inherent in Hegel's philosophical views. At the same time, from a metaphilosophical standpoint, the presence of contradiction is necessary to carry out philosopher proper. Ongoing debate surrounds the question regarding the kind of presence this is.

Hegel's embrace of contradiction also highlights his own conception of Logic. In the *Logic*, Hegel takes it upon himself to re-envision and reformulate the subject:

> As a matter of fact, the need for a reformation of logic has long been felt. In the form and content in which it is found in the textbooks, it must be said that it has fallen into disrepute. It is still being dragged along, more from a feeling that one cannot dispense with a logic altogether and the persisting traditional belief in its importance, than from any conviction that such a commonplace content and the occupation with such empty forms are of any value or use.[23]

---

[19] See Priest 1989, 388 - "A dialetheia is a true contradiction, where "contradiction" has its ordinary, logical sense."
[20] Hegel 2018, 10,S15
[21] See Hegel and Yovel 2005, 90, footnote to The knowing subject
[22] Ibid, 91
[23] Hegel 2010, 31,21.36

He proceeds to claim that Logic has been disfigured and stained with the addition of 'psychological, pedagogical and even physiological' materials. We can see from this part of the preface that Hegel finds the then current status of Logic lacking. However, Hegel's *Logic* is very different in nature from what most people would expect of a book on Logic. There are no symbols and rules, and it starts off with seemingly metaphysical concepts such as "Being", "Nothing" and "Becoming".[24] What then is Hegel's conception of Logic? I believe this question is crucial towards understanding Hegel's embrace of contradiction as part of his approach towards philosophy. In this section of the essay, I would be expounding on two main interpretive approaches towards understanding Hegel's Logic. They are what I have considered as possible ways that Hegel's conception of Logic could have been. Alongside this endeavour, I shall be relating the concept of Logic as "thought thinking itself" to the notion of Hegel's embrace of contradiction.

2.1 Presence of Contradiction – Logic as Metaphysics

One of the paradigms of reading Hegel's *Logic* is metaphysical in nature. (For a very brief historical overview of this approach, see Burbidge (2007, 211-213)) In this respect, Hegel interacts with Plato and Aristotle and exercises a more ontological rather than syntactical or semantical kind of logic. Logical reasoning matters not as much as the positing of 'objects' to account for our understanding as well as the contradiction that arises within our finite human thought. In this case, the *Logic* is seen as a more direct continuation of the agenda in the *PhG*. Paul Rudding relates how Logic manifests as Metaphysics by focusing on the LNC as a manifestation of 'the law of non-compossibility of contraries.[25] The proposition '"It is true and not true" is always false' becomes a matter of whether "individual substances are capable of having incompatible properties at the one time and in the same respect".[26] This idea transforms the LNC into a metaphysical doctrine of which it can then be debated whether or not Hegel actually rejects it.

2.2 Presence of Contradiction - Logic as "thought thinking itself"

Next, Hegel's embrace of contradiction is not his refusal of one objective truth or outcome. Rather, the embrace of contradiction arises from the thought that Philosophy progresses via a logical sequence of "stasis" and "movement". In the *SL,* Hegel outlines how manifestations such as "Being" and "Nothing" move and are united through "Becoming". Hegel's conception of Logic can be interpreted in this case as "thought thinking itself", in other words, the process of our thinking as we come into contact with a particular content and how we then proceed to conceptualise this content.[27] As Burbidge (1993) puts it, 'We are not interested in a casual, psychological dynamic, but rather in the kinds of thinking that are universal and binding, the kinds of thinking most reflective people share. Logic spells out these most basic intellectual operations.'

This conception of Logic can be taken to be continuous and consistent with the deductive inferences of Aristotle as well as the formalisation of such inferences using

[24] This point is adapted from Burbidge "Conception of Logic" 1993
[25] See Redding 2007, 209
[26] Redding 2007, 209
[27] Burbidge 1993, 94

Mathematics a la Russell and Frege. In this case, logic involves the intellectual operation of drawing conclusions from premises and understanding how certain reasoning takes place. In particular, Hegel concentrates on the latter more. When we understand something, we have a particular concept of it. This concept is modified as we pick up new knowledge about it or lose some knowledge that has been unjustified. There is no state of fixity in this process. Once again, the notion of flux as it did for consciousness in the *PhG* re-emerges, but now in relation to logical reasoning and understanding. Our thought processes, in conjunction with our thoughts are in a perpetual state of flux.

To draw on a small example, Hegel begins the *Logic* with "Being"[28] – the most elementary feature of thought. However, despite being a primitive feature of thought, it does not rest easy. In fact, it is unstable. It is only stable when it is mediated through the stage of "Becoming". The unity of "Being" and "Nothing" comes from the realisation that "Being" contains "Nothing" within it. It is not the pure, empty concept it claims to be. The dialectic, in this case, represents the transformative structure or mould that thought brings along as it proceeds from form to form. A parallel analogy with this shell is the phenomenon of arguments in everyday life. For instance, my arguments for the increase in taxes on the rich rests on several sub-arguments, some of which may seem fundamental in nature like the deontic responsibilities that the rich have towards the world. However, my opponent may cast doubts on my main argument by tearing apart my sub-arguments and thereby, dismantle the fundamental position of which my entire argument rests. Therefore, although the notion of back-and-forth arguments may seem like a binary 'yes' or 'no' motion, its underlying features betray a form of Hegelian motion. It is this sense of movement that Hegel's conception of Logic rests on as thought thinks about itself. This movement then characterises Hegel's embrace of contradiction.

Given all the prior grounds that have been set, does Hegel reject the LNC? Philosophers like J.E. McTaggart and Robert Brandom says 'No'. Rather, as per Brandom, Hegel 'radicalises it…and places it at the very centre of his thought.'[29] Therefore, as mentioned a few paragraphs above, Hegel did into call into doubt whether the LNC holds necessarily but does not give a definite answer as to whether it should be rejected. On the other hand, philosophers like Graham Priest uses Hegel as an endorsement of his Dialetheism, which rejects the LNC. In Priest (1989), Priest illustrates how a true contradiction can manifest in the case of differentiating between a body *b* being in motion and being at rest. Hegel's dialectic in this case is directly related to the system that Priest is pushing for in modern logic. However, there are many complications with this view which I would not go through here.

Hegel also views Philosophy as a discipline engaged as much in its development as its end–product. It is for this reason that he claims that a Preface cannot do justice to a piece of philosophical work as the content within it neglects the development of the thesis of the work.[30] I use "thesis" in the sense of a main point or main argument of an expository essay or work, similar to Hegel's Thing (*Sache*) which is "roughly equivalent to '(subject-) matter'…" (Hegel 2018, 330) Even if the development of the thesis consists of contradictions that nonetheless led

[28] For why this is the most basic form of thought to start of our understanding of Logic, see Burbidge 1993, 95 and Hegel 2010

[29] Redding 2007, 201

[30] See Hegel 2018, 5, S1 - "*philosophy moves essentially in the element of universality that embraces the particular within itself*, and this creates the impression, more here than in the case of other sciences, that the Thing itself, in all its essentials, is expressed in the aim and the final results, whereas the elaboration is really the *inessential*." (emphasis mine)

to it, these contradictions should be embraced as part of that philosophical position. Hence, to Hegel, the importance of the development of a philosophical concept or the process of solving a philosophical problem cannot be concluded in the objective doing-away of any stance opposing the final product. This means that as much as there may arise an antithesis of sorts opposing a particular cognitive standpoint, this antithesis does not demolish the standpoint entirely. Therefore, to speak of a standpoint having been refuted would probably be nonsense to Hegel.

3. PHILOSOPHY AS THE DISCIPLINE TO RESOLVE CONTRADICTIONS

This last section is a corollary of the first two points and is hence, not substantive in itself. My motive is just to extract a relation between Hegel and Wittgenstein, so as to conclude that Hegel, similar to Wittgenstein, had a therapeutic view on philosophy. One can say that to Hegel, that is his modus operandi all along in his exposition of consciousness as spirit and the subsequent manifestation within the *Logic*.

Although Hegel thinks that the "truths of philosophy are valueless, and must then be treated as baseless hypotheses, or personal convictions'" without their interconnection with one another within a particular system, he does not eschew philosophy.[31] Instead, Hegel exalts philosophy and held it to be a discipline capable of resolving contradictions. His championing of philosophy can be said to be one of the motivations behind why he pays a large amount of attention to studying human thought itself before the content of thought, so as not to conduct philosophy in a manner that is deemed improper and thereby, to commit a form of double standard. In other words, to do philosophy proper is first to be in a state of Absolute Knowing, as mentioned in section 1. However, although Absolute Knowing is essential to doing philosophy proper, Hegel would not eschew the conduct of philosophy for those who have not achieved it. On the other hand, later philosophers, in particular, Wittgenstein who have a 'Quietist' streak differ in that he seemed to insist that one must abandon philosophy.[32] However, Wittgenstein (whether it is the early or later[33]) and Hegel actually shares a common trait and that is their therapeutic approach in handling philosophical problems.

Philosophy 'is traditionally regarded as a theoretical subject – one that aspires, by more-or-less a priori means, to get to the bottom of things: to unearth the nature of reality, the relations between mind and body, the conditions for knowledge, the right way to conduct one's life, and so on.'[34] It is a mainly an *a priori* subject that operates by human thought which, in turn is known or expressed through language. It is for this reason that analytic philosophers, especially in the turn of the 19th century and upon the realisation of philosophy's dependence on language, focused more on the structure of language.[35] Wittgenstein – the later Wittgenstein – was one such figure who came to the conclusion that since language could not be divorced

[31] See Stern 2013

[32] Wittgenstein ends the *Tractatus-Logico-Philosophicus* by claiming that "What we cannot speak about, we must pass over in silence"(L. Wittgenstein 2009b) stressing the point that philosophical problems are merely present due to our flawed views of language, especially since we fail to understand the logical structure and relationship that language has with the world. Hence, we should take a non-interventionist stance in philosophy i.e. abandon philosophy in the sense of not making any substantive, positive contributions. See McDowell 2009

[33] See Horwich 2020

[34] See Horwich 2013

[35] See Carnap 1932; Frege 1960; Russell 1905

from its use, hence there would be no end to the search for the logical structure of language as meaning is simply not bound by truth conditions. This would also imply that to do philosophy in the manner of philosophical theorising or conceptualising would be flawed. As a result, one should give up philosophy in the manner mentioned. What then becomes of the task of philosophers? Wittgenstein thinks that what remains for philosophers is not to tackle problems or resolve contradictions head on but to uncover the false foundation on which the problem lies and hence, to 'dissolve' the issues.[36] Meanwhile for Hegel, he thinks that it is only with 'further philosophical reflections that we can see our way through the problems…'[37] Therefore, while philosophical problems are treated as pseudo-problems by both philosophers, Hegel, in particular, utilises the tools of philosophy i.e. *a priori* thinking to 'dissolve' these problems, rather than redefine the agenda of philosophy, as Wittgenstein does.

CONCLUSION

A towering figure who can be said to have reigned the throne of German Idealism before the takeover by positivist and empiricist Analytic philosophers near the 20th century, Hegel was a genuine intellectual figure, albeit a polarising one. His systematic philosophy was certainly special but not one that is easily acceptable. In this paper, I first elucidate Hegel's conception of philosophy as a domain of human thought, elaborating on his insistence on philosophy's *a priori* nature as opposed to the empirical sciences' treatment of the *a posteriori* domain. Subsequently. I talked about how Hegel embraced contradiction, rather than tried to eliminate them head-on. In this respect, Hegel can be said to have represented one of the first few Western philosophers to have either advocated for a denunciation of the LNC or at least, expressed a form of scepticism about it. I then conclude that Hegel's embrace of contradiction and his viewing of philosophy as a primarily cognitive domain led him to embrace also a form of therapeutic approach that he shares with philosophers like Wittgenstein.

REFERENCES

Berlin, Isaiah. 2013. *The Hedgehog and the Fox: An Essay on Tolstoy's View of History*. Edited by Henry Hardy. Second ed. Vol. Book, Whole: Princeton University Press.

Burbidge. 1993. *Hegel's conception of logic* Edited by Frederick C. Beiser. Vol. Book, Whole*The Cambridge companion to Hegel*: Cambridge University Press.

———. 2007. "The relevance of Hegel's Logic." *Cosmos and history* 3 (2-3): 211-221.

Carnap, Rudolf. 1932. "The Elimination of Metaphysics Through Logical Analysis of Language." *Erkenntnis*: 60-81.

Forster, Michael N. 2019. "The Origin and Character of Hegel's Concept of Geist." 29-54. Cambridge University Press.

[36] In the words of Wittgenstein, his aim in philosophy was to show the 'fly the way out of the fly bottle' Wittgenstein 2009a, 110

[37] Stern 2013, 17

Frege, Gottlob. 1960. "On sense and reference." In *Arguing About Language*, edited by Darragh Byrne and Max Kölbel, 36--56. Routledge.

Giovanni, George di, and H.S. Harris. 1985. *Between Kant and Hegel: Texts in the Development of Post-Kantian Idealism*. State University of New York Press.

Hegel, Georg Wilhelm Friedrich. 2010. *The Science of Logic*. Translated by George Di Giovanni. Vol. Book, Whole: Cambridge University Press.

———. 2018. *Hegel: The Phenomenology of Spirit*. Translated by Michael J Inwood. First ed. Vol. Book, Whole: Oxford University Press.

Hegel, Georg Wilhelm Friedrich, and Yirmiyahu Yovel. 2005. *Hegel's preface to the Phenomenology of spirit*. Vol. Book, Whole. Princeton, N.J: Princeton University Press.

Horwich, Paul. 2013. "Reply to Timothy Williamson's Review ofWittgenstein's Metaphilosophy: Reviews." *European journal of philosophy* 21: e18-e26.

———. 2020. "Wittgenstein on Truth." In *WITTGENSTEINIAN (adj.): Looking at the World from the Viewpoint of Wittgenstein's Philosophy*, edited by Shyam Wuppuluri and Newton da Costa, 151-162. Springer International Publishing.

Kojeve, Alexandre. 1980. *Introduction to the reading of Hegel*. Translated by Raymond Queneau. Vol. Book, Whole: Cornell University Press.

McDowell, John. 2009. "Wittgensteinian "QUIETISM"." *Common knowledge (New York, N.Y.)* 15 (3): 365-372.

———. 2018. *What is the Phenomenology about?* Edited by Federico Sanguinetti, André J. Abath and SpringerLink. Vol. Book, Whole*McDowell and Hegel: Perceptual Experience, Thought and Action*: Springer International Publishing.

Pippin, Robert. 1993. "You can't get there from here: transition problems in Hegel's Phenomenology of Spirit." In *The Cambridge Companion to Hegel*, edited by Frederick C. Beiser, 52--85. Cambridge University Press.

Priest, Graham. 1989. "Dialectic and Dialetheic." *Science & Society* 53 (4): 388-415.

Redding, Paul. 2007. *Hegel and Contradiction*. Vol. Book, Whole*Analytic philosophy and the return of Hegelian thought*: Cambridge University Press.

Russell, Bertrand. 1905. "On Denoting." *Mind* 14 (56): 479-493.

Smith, Robin. 2020. *Aristotle's Logic*. Fall 2020 ed.*The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta: Metaphysics Research Lab, Stanford University.

Stern, Robert. 2013. *The Routledge Guide Book to Hegel's Phenomenology of Spirit*. Routledge.

Taylor, Charles. 1977. *Hegel*. Vol. Book, Whole: Cambridge University Press.

Wittgenstein. 2009a. *Philosophical investigations*. Translated by G. E. M. Anscombe, P. M. S. Hacker and Joachim Schulte. Rev. 4th ed. Vol. Book, Whole: Wiley-Blackwell.

———. 2009b. *Major Works: Selected Philosophical Writings*. HarperCollins.

REVIEWS

MONTAIGNE: A VERY SHORT INTRODUCTION
WILLIAM M. HAMLIN
Oxford, Oxford University Press, 2020, xxi + 140 pp., bibliography and index

MICHEL DE MONTAIGNE was born in 1533 in Aquitaine on the family estate Château de Montaigne. Born an aristocrat into a wealthy family – as a result of his grandfather, Ramon Felipe Eyquem's mercantile ventures which allowed him to purchase the estate in 1477 – Montaigne was educated, somewhat like the now-more-famous example of John Stuart Mill, according to his father's pedagogical plan. He learnt Latin as a first language and was over six before he knew 'any more French or Perigordian than Arabic'.[1] In 1539, he was sent to study at the prestigious College of Guienne, where he mastered the curriculum by thirteen. He went on to study law and became a councillor of the Court des Aides of Périgueux and, in 1557, of the Parlement in Bordeaux. From 1561 to 1563, he was courtier to Charles IX. It was whilst at the Bordeaux Parlement that he became close friends with Étienne de La Boétie, whose early death in 1563 caused him acute sadness even eighteen years later. He wrote in his *Travel Journal* that 'no spoken or written statement in the schools of philosophy ever represented the rights and duties of sacred friendship as exactly as did the practice that my friend and I formed together'.[2] He married Françoise de La Chassaigne in 1565, who bore six daughters from 1570 to 1583, but only one, Léonor, born in 1571, lived beyond infancy. In 1570, he sold his councillorship and isolated himself in the Tour de Montaigne for a period of ten years, during which time he wrote his *Essays*, which were published in 1580. He began to suffer from kidney stones in 1578, and it was partly in search of a cure that he travelled to France, Germany, Austria, Switzerland, and Italy in 1580 and 1581; but it was also a pilgrimage to the Holy House of Loreto. It was during this time that he wrote his *Travel Journal* which, though it was never intended for publication, was published nevertheless in 1774. He returned to France in 1581 after discovering that he had been elected mayor of Bordeaux, and remained there until 1588, when he travelled to Paris to arrange the publication of a new edition of the *Essays*. He died of disease in 1592.

It was in his decade long sojourn in the Tour de Montaigne, with its library furnished with some fifteen hundred works, that he wrote the first edition of his *Essays*, which were published in 1580, and are now know to us as the A passages. He continued to work on them, however, and published a revised edition in 1582 with several additions reflecting the concerns of Vatican censors, which are known as the A1 passages. His journey to Paris resulted in the publication of a third edition in the same year, which added the Third Book of the *Essays* and six hundred additions to the first two books, which are known as the B passages, and of which the famous Bordeaux Copy is exemplar. Finally, he made further editions from 1588 until his death in 1592, which were published in 1595 and are known as the C passages. In 1603, the *Essays* were translated into English by John Florio, who relied on the final edition printed in 1595.

The *Essays* were, according to Montaigne, 'the only book in the world of its kind, with a wild and eccentric plan'.[3] It is perhaps because of he himself that there is the misconception

that Montaigne was the progenitor of essayists; really, there are a great many works that we call essays to be found in antiquity – which Montaigne refers to himself in his *Essays* – and his essays are at many times not essays at all, but what we might consider either confessions or idle thoughts (*zuihitsu* 随筆). Regarding the former, he says: 'In honour of the Huguenots, who condemn our private confession, I confess myself in public, religiously and purely… I am hungry to make myself known, and I care not to how many, provided it be truly'.[4] And of the latter: 'These are my humours and opinions; I offer them as what I believe, not what is to be believed'.[5] Obviously, Augustine's *Confessions* and Yoshida Kenkō's (兼好) *Essays in Idleness* (*Tsurezuregusa* 徒然草) predate the *Essays* considerably too.

It is on the *Essays* that William M. Hamlin's *Montaigne: A Very Short Introduction* focusses. In the preface he writes:

> This book is intended for a general audience, and particularly for those encountering Montaigne for the first time. I have sought to make it accessible to upper-level undergraduates, to graduate students, and to general readers looking for a broad introduction to the life and thought of the essayist.[6]

There is, however, only a single chapter on Montaigne's life, with the rest of the book organized conceptually around his opinions on certain ideas. Mr. Hamlin treats Montaigne as a philosopher, therefore, and not merely an essayist. As well as the ideas within the *Essays*, Mr. Hamlin details their construction and, of greater importance, their reception, which is academically excellent, for it is proper that one should consider in depth not only contemporary assessments of a philosophy, but how it was received in a different time. This is quite important, because contemporary thought is the product of circumstance – which is why Orwell wrote about Communism, Dostoyevsky about suffering and Wollstonecraft about suffrage – and circumstance is the product of previous thought – which is why the Communists spoke often of Marx, I speak about Dostoyevsky, and feminists, at least those interested in equality rather than supremacy, speak of Wollstonecraft. And indeed, Montaigne thought that his *Essays* would be socially useful: 'What is useful to me might also by accident be useful to another'.[7]

Montaigne is not greatly known for his pedagogy – it is worthy of note that he is not included within Routledge's fifty major thinkers on education[8] – but Mr. Hamlin reveals a forward-thinking, highly philosophic pedagogy within the *Essays* that is remarkable at a time prior to the Enlightenment when religious dogma still ran rife. Montaigne believed in absolute truths but thought that the greatest victories were to be found in conversation, debate and changes of mind:

> If I had had to train children, I would have filled their mouths with this way of answering, inquiring, not decisive—'What does that mean?' 'I do not understand it.' 'That might be.' 'Is it true?'—so that they would be more likely to have kept the manner of learners at sixty than to represent learned doctors at ten.[9]

Both interlocutors, however, must be prepared to engage in this dialectic: 'It is impossible to discuss things in good faith with a fool'.[10] In this sense, he is certainly worthy of consideration as a philosopher; we see here the dialectical method, of which there is seldom anything more representative of philosophy in the West, working under the surface of the *Essays*. Montaigne believed in an 'essential pattern' (*forme maistresse*) that was the core of selfhood, not immune to change but sufficiently secure for long-term traits to form: 'There is no one who, if he listens

to himself, does not discover in himself a pattern all his own, a ruling pattern, which struggles against education and against the tempest of the passions that oppose it'.[11] He did not think that this essential pattern rendered education useless, but saw the limits of pedagogy. About this, Mr Hamlin writes:

> Relying in part on the Platonic theory that all earthly phenomena may be understood as imperfect manifestations of original and perfect "forms," Montaigne's sense of this *forme maistresse* is not that of a fixed identity but of an inborn dispositional frame within which specific identity-formations can develop and mature. This frame is neither immutable nor impervious to external influence, but on the whole its boundaries are firm enough to encourage certain long-term tendencies while discouraging others.[12]

As for the role of teachers, I am inclined to agree with Montaigne:

> Let the tutor make his pupil pass everything through a sieve and lodge nothing in his head on mere authority and trust: let not Aristotle's principles be principles to him any more than those of the Stoics or Epicureans. Let this variety of ideas be set before him; he will choose if he can; if not, he will remain in doubt.[13]

It might be remarked that a great many professors ought to consider carefully this philosophy, for it is rare that a rationalist does not teach other rationalists, an analytic professor does not teach analytic students, and those dedicated to a certain literary style do not insist that there is only one way to write.

In its Renaissance usage, friendship (*amitié*) denoted a wider range of relations than in the contemporary common parlance, and love (*amour*) was typically reserved for erotic relationships between men and women. Of these affections, Montaigne says:

> To compare this brotherly affection with affection for women, even though it is the result of our choice—it cannot be done; nor can we put the love of women in the same category. Its ardor, I confess, … is more active, more scorching, and more intense. But it is an impetuous and fickle flame, undulating and variable, a fever flame, that holds us only by one corner. In friendship [*amitié*] it is a general and universal warmth, moderate and even, besides, a constant and settled warmth, all gentleness and smoothness, with nothing bitter and stinging about it. What is more, in love [*amour*] there is nothing but a frantic desire for what flees from us.[14]

He draws this theory from 'four ancient types' of affection: 'natural, social, hospitable, erotic'.[15] This largely reflect the classical Graeco-Christian taxonomy, which classified love into familial (*storgē* στοργή), brotherly (*philia* φιλία), neighbourly (*agapē* ἀγάπη) and erotic (*érōs* ἔρως). For Montaigne, 'we do not marry for ourselves, whatever we say; we marry as much or more for our posterity, for our family',[16] and 'few men have married their mistresses who have not repented it'.[17]

It was not just in education that Montaigne was progressive: he also had some recognition of what we today call introverted and extroverted personalities: 'there are private, retiring, and inward natures',[18] but he is himself 'all in the open and in full view, born for company and friendship. The solitude that I love and preach is primarily nothing but leading my feelings and thoughts back to myself, … I throw myself into affairs of state and into the world more readily when I am alone'.[19] 'He is, in short, an extrovert who seeks solitude in

society and society in solitude'.[20] As for this solitude, it is good in small doses, restoring us to ourselves and improving our perception for our return to society. This makes it important, because, according to Montaigne, we have a duty to be fully ourselves, and improving our self-knowledge and self-reliance could be an effective defence against loss or even death.

In keeping with his admiration of solitude, Montaigne's main concerns are personal independence and liberty. He says: 'I am so sick for freedom, that if anyone should forbid me access to some corner of the Indies, I should live distinctly less comfortably'.[21] He is, nevertheless, acutely aware of how mortality and society encroach on personal liberty. 'A list of such infringements would include, at the very minimum, age, chance, habit, family circumstance, state of bodily health, local custom, individual temperament, cultural expectation, zealous commitment to a cause, and adherence to a particular metaphysical system' (59-60).[22] Less consistent with his thinking on solitude and liberty, however, is his conclusions that social existence inevitably forces us to subordinate our individual desires to those of the group.[23] He was, ultimately, a conformist, partly as a result of his conservativism, partly because of a belief in a divine providence which infallibly placed everything such that, as Professor Pangloss says, 'everything is for the best' (*tout est pour le mieux*).[24]

Montaignian forward-thinking extends from pedagogy and psychology to intercultural confrontation and diversity too. Whereas his contemporaries in the Old World looked upon the native Americans as savages, Montaigne thought that the people of the New World, though they are not in a state of prelapsarian innocence, represented a human existence closer to 'original naturalness' than that of the Europeans.[25] 'Montaigne interprets the very fact of difference between the New World and the Old as a tension between natural and artificial behavior'.[26] He sees virtue as a spectrum, with natural virtue as less evolved and the overcoming of natural vice to become artificially virtuous as more valuable. He abhorred cruelty and mourned the razing of the New World for 'pearls and pepper'.[27] The peoples of the Americas were, for him, a useful rhetorical device for contrasting their 'savagery', such as their cannibalism, with the calculated cruelties of the Europeans.

As for diversity, Montaigne was so avid as many today, but his motivation was philosophic rather than essentialist. By diversity he means intellectual and cultural diversity, rather than the sexual and racial diversity which we find insisted on today; the latter necessarily presupposes that people of a certain race or sex are 'like that', and that it is impossible to look alike and think differently – hence nobody calls a group of white men today diverse, but a group of various races who all think alike is considered a paragon to aspire to – but Montaigne did not fall into such essentialist dogmatism. Whether this has anything to do with his coming before the advent of critical race theory and postcolonial discourse, and the advent of this essentialist malaise coinciding with it, that perhaps the two are related, it is best left to the imagination, least of all because such essentialists are generally incapable of distinguishing between philosophic and political matters. To digress, he says that 'the world is nothing but variety and dissimilarity',[28] and 'nature has committed herself to make nothing separate that is not different'.[29] He warns against the human tendency to judge the unfamiliar as inferior: 'It is a common vice, not of the vulgar only but of almost all men, to fix their aim and limit by the ways to which they were born'.[30] He did not advocate childishness, but recognition that our thoughts could be as flexible as a child's, and that we must assess the monstrous and miraculous in terms of human finitude, rather than in terms of our cultural dogma: 'We must not judge what is possible, and what is not, according to what is credible and incredible to our sense'.[31]

Philosophically, 'no characterization of Montaigne has held greater sway than that he is a skeptic'.[32] He had read Sextus Empiricus' *Outlines of Pyrrhonism* a few years after he began drafting his *Essays*, and his philosophy came to synthesize Academic and Pyrrhonian Skepticism. He thinks that we are frail and imperfect and in many respects inferior to animals. He seeks to undermine religious dogmatism (but not the official religious dogma), especially as regards Raymond Sebond's *Natural Theology*, which Montaigne had translated from Latin into French but had been placed on the Vatican's *Index of Prohibited Books*. He concludes that, even if we can acquire knowledge through our own faculties, which is, he thinks, dubious, such knowledge is vain unless infused with divine grace. In Montaignian Skepticism, the ideal teacher exposes students to multiple perspectives rather than dictating their beliefs. Ignorance is his 'ruling quality',[33] and 'all the abuses of this world are engendered by our being taught to be afraid of professing our ignorance and our being bound to accept everything that we cannot refute'.[34]

> What we might call "Montaignian skepticism" is thus a pragmatic and case-oriented synthesis of doubt, inquiry, and provisional conclusion. It embraces ignorance, it valorizes detached investigation, and it prizes humility and self-critique, but it also assumes that certain truths exist. It is acutely sensitive to linguistic nuance, as when Montaigne tells us that "I love those words which soften and moderate the rashness of our propositions: 'perhaps,' 'to some extent,' 'some,' 'they say,' 'I think,' and the like."[35] At times it takes the form of relativism, and Montaigne is justly famous for imagining the world from a feline perspective: "When I play with my cat, who knows if I am not a pastime to her more than she is to me?".[36] And it has a particularly vexed relationship with the power of authority.[37]

More renowned than even his skepticism, though it is more famous as an aphorism – especially in reference to the death of Socrates – than a Montaignian quote, is "That to Philosophize Is to Learn to Die," which is one of his essays. For Montaigne, 'it is uncertain where death awaits us… let us await it everywhere. Premeditation of death is premeditation of freedom. He who has learned how to die has unlearned how to be a slave. Knowing how to die frees us from all subjection and constraint'.[38] Death forms part of any meaningful life:

> In everything else there may be sham: the fine reasonings of philosophy may be a mere pose in us; or else our trials, by not testing us to the quick, give us a chance to keep our faces always composed. But in the last scene, between death and ourselves, there is no more pretending; we must talk plain French, we must show what there is that is good and clean at the bottom of the pot.[39]

He is optimistic in his early essays about the emancipation of life by death, but is later less confident that we can detach ourselves from the constraints of mortal existence. He is convinced that, either way, death is firmly bound with life – which is true, hence his philosophy of mortality is all the more important.

Mr. Hamlin's *Montaigne* is conceptually organized like this, and therefore has the hallmark of a weak book, potentially being more of an introductory selection than a proper introduction; but his selection is less selective than it is comprehensive, his having taken from the *Essays* the thoughts of philosophic import, and leaving behind those that are more rudimentary or diarylike. The book is also excellently referenced. Mr. Hamlin quotes liberally from the *Essays* and includes the sources for other comments found in the book. It is

unfortunate, in fact, that some of Oxford University Press' other very short introductions do not resemble Mr. Hamlin's. They often make terrible use of the space afforded them, focussing horribly on tiring explanations, analogies and examples rather than the concise and direct summary that one would expect of an introduction so short, and rely heavily on certain interpretations or texts which are never cited. Mr. Hamlin's book is an exemplar in this respect. Perhaps a single reservation would be on aesthetic terms, that he quotes from the edition translated by Donald M. Frame (1957-1958), which is in Modern English, whereas that by John Florio (1603) or by Charles Cotton (1685-1686), though it was later edited by William Carew Hazlitt (1877), are is Archaic and Obsolete English, which might be imagined as being closer to the Middle French in which the *Essays* were originally written. This criticism, however, is of little concern: it is common and perfectly academic to substitute an older translation for a newer one, and there are, as always, advantages and disadvantages to both.

B.V.E. HYDE
Durham University

[1] Michel de Montaigne, *Essays*, I.26.128.
[2] Michel de Montaigne, *Travel Journal*, 1060.
[3] Michel de Montaigne, *Essays*, II.8.278.
[4] Ibid. III.5.643.
[5] Ibid. I.26.108.
[6] William M. Hamlin, *Montaigne: A Very Short Introduction*, page xx.
[7] Michel de Montaigne, *Essays*, II.6.272.
[8] Palmer, Joy (Ed.), (2001), *Fifty Major Thinkers on Education: From Confucius to Dewey*, London: Routledge.
[9] Michel de Montaigne, *Essays*, III.11.788.
[10] Ibid. III.8.706.
[11] Ibid. III.2.615.
[12] William M. Hamlin, op. cit., page 41.
[13] Michel de Montaigne, *Essays*, I.26.111.
[14] Ibid. I.28.137.
[15] Ibid. I.28.136.
[16] Ibid. III.5.645-46.
[17] Ibid. III.5.649.
[18] Ibid. III.3.625.
[19] Ibid. III.3.625
[20] William M. Hamlin, op. cit., page 57.
[21] Michel de Montaigne, *Essays*, III.13.820.
[22] William M. Hamlin, op. cit., pages 59-60.
[23] Michel de Montaigne, *Essays*, III.5.648, III.9.758.
[24] Voltaire, *Candide*, *passim*.
[25] Michel de Montaigne, *Essays*, I.31.153.
[26] William M. Hamlin, op. cit., page 66.
[27] Michel de Montaigne, *Essays*, III.6.695.
[28] Ibid. II.2.244.
[29] Ibid. III.13.815.
[30] Ibid. I.49.215.
[31] Ibid. II.32.548.
[32] William M. Hamlin, op. cit., page 84.
[33] Michel de Montaigne, *Essays*, I.50.219.
[34] Ibid. III.11.788.
[35] Ibid. III.11.788.
[36] Ibid. II.12.331.
[37] William M. Hamlin, op. cit., pages 94-95.
[38] Michel de Montaigne, *Essays*, I.20.60.
[39] Ibid. I.19.55.

FRANK RAMSEY: A SHEER EXCESS OF POWERS
CHERYL MISAK
Oxford, Oxford University Press, 2020, 544 pp., bibliography and index

A PHILOSOPHER IS ONLY sincere when he lives as his philosophy commends, or, conversely, when his philosophy reflects who he truly is. By this criterion, Frank Ramsey is probably one of the sincerest philosophers ever lived. His pragmatic and humanistic philosophy is, to a great extent, a beautiful reflection of his vigorous and enthusiastic personality; consequently, some knowledge of the latter will corroborate an understanding of the former. This is what Cheryl Misak's brilliant biography of Ramsey enables us to do: it delivers a delicate balance between the intellectual and the personal aspect of Ramsey's life, situates his philosophical works in the context of his life, and thus provides us with a profound understanding of what Ramsey's pragmatist philosophy is about.

The book is divided into three parts. The first deals with Ramsey's childhood, and introduces the intellectual background in Cambridge at the time. The second recounts his undergraduate life, his impressive presence as a young prodigy in Cambridge, and his emotional struggles during the period. The third focuses on his astonishing achievements in philosophy, mathematics and economics during the half decade he had before his death.

Misak in general deals with these incredibly wide-ranging academic topics at ease and makes them highly accessible to the readers without much professional knowledge in the area. Wherever she feels inadequate, she resorts to relevant experts in the field and asks them to write a page-long exposition of the topic which she then puts in the 'boxes' of her book. The boxes, however, are less impressive than the other contents. They are too short to supply much information for both the layman and the professional: the former will find it too compressed and the latter may find it too cursive. Other than that, the book is a brilliant masterpiece. It combines width and depth, the personal and the academic, and is written in a suitably lucid and concise manner. This review provides some background to the book and knits some key points in it so that its gist makes better sense to the readers.

Ramsey is, without much doubt, one of the greatest pioneers of pragmatism. It is therefore useful to first explain what pragmatism is. Following Daniel Williams,[1] I see pragmatism as consisting of the following three pillars:

1. Primacy of the practical,[2] or the replacement of 'copying' by 'coping'.[3] According to pragmatists, what we do is prior to what we say, and saying is a kind of doing. By uttering sentences in a language, we do not attempt to represent the world 'as it is anyway';[4] rather, language is primarily a special kind of act, in order to communicate, cooperate, and cope with the vicissitudes of the world. This is pragmatism as anti-representationalism (championed by Richard Rorty): language is not a mirror that reflects the world;[5] rather, it is (according to the Later Wittgenstein) a toolbox that we use to serve our various practical ends.[6]

2. Human contingency. Pragmatists deeply recognise the contingency of our own perceptive and cognitive faculties, and encourage us not to project our own productions onto the world (common examples: morality, causation, induction, time…). As William James famously puts it, 'the trail of the human serpent is… over everything'.[7] This is echoed by contemporary

pragmatists like Huw Price, whose pragmatism as global expressivism seeks an understanding of our language and vocabularies in terms of our 'contingent, shared dispositions' and 'practical stances'.[8]

3. Social pragmatism about normativity.[9] For pragmatists, all normativity is ultimately derived from social facts. That is to say, there is nothing divine above us, nothing that we should be responsible for other than our fellow human beings. The more obvious example is moral phenomenon, whose normativity is more commonly seen as social. The less obvious examples are intentionality and meaning: for pragmatists, both are normative (one can believe or speak *rightly or wrongly*), and therefore both are social. A certain term means what it does (which implies that we *should* apply it in this way) simply because we – the language community – *do* use it in a certain way, and there is nothing over and beyond this fact that could endow anything with normativity.

I see these three pillars as being underpinned by the common theme of humanism. It is humans that have all kinds of practical ends, that have contingent faculties and contingent languages, and that are essentially social beings. While the 'metaphysical' philosophy concerns itself with objective reality, absolute truth and external facts, the pragmatist philosophy focuses on the human perspective, human ends and human choices. Ramsey is a pragmatist insofar as the emphasis of human perspective and the concern for human wellbeing lie at the centre of his philosophy. As he himself puts it: 'My picture of the world is drawn in perspective, and not like a model to scale. *The foreground is occupied by human beings* and the stars are all as small as threepenny bits' (my emphasis).[10]

There are multiple examples, throughout Ramsey's unfortunately short life, that can testify to this claim. In very early stages of his academic career (when he was still an undergraduate), he argued against John Maynard Keynes's account of probability and justification of inductive inference, on the ground that both probabilistic and inductive inference are *psychological* rather than objective.[11] For Keynes, probability is an objective relation between any set of premises and a conclusion, something we can directly perceive and cannot be further analysed. Ramsey, on the other hand, is suspicious of any property that is both objective and unanalysable; he reverses the order of Keynes's explanation, and argued (more extensively in his later "Truth and Probability") that probabilistic knowledge is to be understood in terms of *subjective degrees of belief*, which were to be measured by one's willingness to bet. Similarly, contrary to Keynes's attempt to underpin induction by his hypothesis of limited variety of properties in nature, Ramsey argues that we should believe in induction simply because it is a *good habit* and it works, and despite that we cannot give any non-circular justification for induction, 'In this circle lies nothing vicious'.[12] This is surely a kind of pragmatism: for him, it is our psychology and our habits, instead of properties of nature, that underlie our various *practices of inference*.

Another telling example is Ramsey's treatment of truth and propositions. He famously proposes a deflationary account of truth, which claims that all we need to know about truth consists of the following platitude: 'a belief that p is true iff p' (this is vulnerable to the Frege-Geach point that truth predicates can be embedded in complex sentences. However, Ramsey's insight inspires modern pro-sentential and minimalist theories of truth which survive and

thrive). This simple analysis, however, only constitutes a small part of the analysis of truth; the main heavy-lifting work is to be done by asking what it is for a belief to be a belief *that p*. For Ramsey, this is to be spelt out in functionalist terms, in terms of its place in the complex web of causes and effects,[13] of what tends to produce the belief and what the belief tends to produce, and thus of what the belief is disposed to *do for us*. This highly original pragmatist analysis of truth and meaning helps to lift the mysterious veil in front of these two concepts, and inspires later (broadly speaking, pragmatic) philosophers to develop more sophisticated accounts along these lines (such as Putnam's functionalism,[14] Millikan's biological version of success semantics,[15] Horwich's[16] and Price's[17] minimalism, etc.).

Apart from the more theoretical endeavours, Ramsey's pragmatism is also reflected in his enthusiasm and involvement in political movements. During his undergraduate times, he actively participated in the post-war socialist movements (as a member of the Cambridge University Socialist Society) and became increasingly concerned with the welfare of the working class. His later works in economics (tax and savings) and his academic engagement with Keynes and Pigou are, to a great extent, motivated by concerns of social justice and utility optimisation. And in political and moral philosophy, he encourages us to always focus on the realistic questions ('What is the world like? How to make it better?') instead of abstract ones ('the so-called paradox of self-government'), as the latter could only lead us to 'fairy tales' instead of truths.[18]

As Misak notes and stresses throughout her book, Ramsey's pragmatic philosophy which places humanity at its centre, at least in part, stems from his vigorous and passionate personality. The general point is powerfully illustrated by Fichte: 'a philosophical system is not a dead piece of furniture that we can reject or accept as we wish; it is rather a thing animated by the soul of the person who holds it.'[19] The 'soul' of Ramsey, as Keynes describes in his obituary, has a boyish enthusiasm, a 'spontaneous gurgling laugh', an honesty of mind and heart, and a relentless curiosity towards all sorts of knowledge (philosophy, mathematics, logic, economics, politics…).[20] Naturally, this kind of personality gives rise to a vigorous philosophy centred around humanity and a concern for its wellbeing. This is manifested most clearly in one of Ramsey's most famous passages:

> Humanity, which fills the foreground of my picture, I find interesting and on the whole admirable. I find, just now at least, the world a pleasant and exciting place. You may find it depressing; I am sorry for you, and you despise me. But I have reason and you have none; you would only have a reason for despising me if your feeling corresponded to the fact in a way that mine didn't. But neither can correspond to the fact. The fact is not in itself good or bad; it is just that it thrills me but depresses you.[10]

This paragraph can be further illustrated by putting Ramsey in contrast with Wittgenstein, even though they eventually get to similar places. The Earlier Wittgenstein is 'the king of representationalism', as Robert Brandom puts it:[21] in *Tractatus* he gives the most elaborate account of how languages and thoughts represent the world. The Later Wittgenstein, to the contrary, is 'the king of anti-representationalism': in *Philosophical Investigations* he entirely refutes his previous work and develops a highly pragmatic account of language, truth, and the world (which is close to, but more sophisticated than, Ramsey's view on these matter; it is only natural as Ramsey died so young). Misak argues, in this book, that Wittgenstein's subversive change is to some extent due to the influence of Ramsey. In 1929 Wittgenstein returned to Cambridge and lived with Ramsey for a while, and during the time they exchanged

a lot on their philosophical views. As Misak documents in detail, Ramsey's critique of the Tractarian project during this time was a substantial factor which led Wittgenstein to see his previous project as indefensible and come along to the pragmatist side.

However, despite all this, they still differ in their philosophical temperaments, rooted in their different personalities. Wittgenstein is constantly unhappy or even painful. He recommends 'serious thinking', reverence, and purity. He, as Frances Marshall describes, does not find the world his friend. He wants to get to the bottom of 'essence' of all things. Ramsey, on the other hand, is constantly smiling and delightful, light and irreverent; he sees Wittgenstein's problem as profound ones, but decides to answer them on a human scale nonetheless. When they eventually reached a similar place, Wittgenstein anguishes and bemoans, while Ramsey remains cheerful about it. Wittgenstein finds it a depressing truth that our rule-following behaviour might not be rationally justified and merely 'something animal'; Ramsey, however, delights in precisely this fact, in seeing nothing divine and essential, in being able to live lightly and pragmatically. As Ramsey wrote, foreshadowing his later conversation with Wittgenstein: 'The fact is not in itself good or bad; it is just that it thrills me but depresses you.'

And finally, it is useful to place Misak's new masterpiece in more context in order to better understand it. Misak is a pragmatist herself, and indeed one of the most famous contemporary pragmatists. As Brandom comments, no one has done more to transform and improve our understanding of the tradition of pragmatism than what Misak has done (and is doing).[22] In her 2013 book, *The American Pragmatists*,[23] she distinguishes two substantially distinct strands of American pragmatism: one runs from C.S. Peirce to C.I. Lewis and then to mid-20th century analytic philosophers like Wilfried Sellars and W.V.O. Quine. The other runs from William James to John Dewey, and finally to Richard Rorty. She recommends the first, the 'rationalist' strand, emphasising the importance of logic and rigorous thinking, and sees the second – the 'romantic' strand – as the regressive wing of pragmatism, commending literature and art as superior to science and logic.

Then, in her 2018 book, *Cambridge Pragmatism: From Peirce and James to Ramsey and Wittgenstein*,[24] she quite originally identifies a line of influence from Peirce to Ramsey, and Ramsey to Wittgenstein. Approaching the end of his undergraduate days, Ramsey came to read some works by James and Peirce. He, like Russell, did not think much of James, but he thought that a lot can be learnt from Peirce's pragmatist perspective (especially his account of belief). He recorded his understanding and approval of Peirce in his diary in 1924, and acknowledged Peirce's influence explicitly in his later paper "Truth and Probability". Misak accurately identifies the influence, as well as Ramsey's influence on Wittgenstein, and thereby separating a strand of pragmatism which she calls Cambridge pragmatism, represented by Ramsey and Wittgenstein in the early 20th century, and later taken up by Simon Blackburn and Price in the beginning of 21st century.

Her new work on Ramsey, as I see it, can be best understood as a continuation of her effort in identifying and synthesising the rich pragmatist tradition. Pragmatism is not only a philosophical theory; it is a way of thinking about the world and ourselves. It is a meta-level framework that changes how we see and live our lives. This is a point that, in its nature, cannot be made in a purely theoretical way; it has to be illustrated with convincing examples, and this is what Misak's previous works – and in general, the previous literature on pragmatism – lacked. Ramsey is the perfect example that Misak eventually found. By telling a story of his life and philosophy, Misak not only exposes how pragmatism is closely connected to life, but

also establishes an epitome of a true pragmatist, one who is often smiling and delightful, honest and open, who has a profound love for life and for fellow human beings, who focuses on the practical and the realistic rather than the abstract, who has a relentless curiosity and enthusiasm towards all kinds of experiences and knowledge, and who, in Joseph Schumpeter's memorable words, has a 'sheer excess of powers'.[25]

BOJIN ZHU
The University of Cambridge

---

[1] Williams, Daniel, (2018). "Pragmatism and the predictive mind". *Phenomenology and the Cognitive Sciences* 17(5): 835-859.
[2] Brandom, Robert, (1994). *Making it Explicit*. Harvard University Press.
[3] Rorty, Richard, (1989). *Contingency, Irony, and Solidarity*. Cambridge University Press.
[4] Williams, Bernard, (1986). *Ethics and the Limits of Philosophy*. Routledge.
[5] Rorty, Richard, (1979). *Philosophy and the Mirror of Nature*. Princeton University Press.
[6] Wittgenstein, Ludwig, (1953). *Philosophical Investigations*. Blackwell.
[7] James, Williams, (2000). *Pragmatism and Other Writings*. Penguin Books.
[8] Price, Huw, (2011). *Naturalism without Mirrors*. Oxford University Press.
[9] A term copied from Robert Brandom's lectures on pragmatism in 2020. See https://www.pitt.edu/~rbrandom/.
[10] Ramsey, Frank, (1925). "On There Being No Discussable Subject". Published posthumously under the title "Epilogue" in Ramsey (1931), *The Foundations of Mathematics and Other Logical Essays.* Ed. R.B. Braithwaite. Routledge and Kegan Paul.
[11] Ramsey, Frank, (1922). "Mr. Keynes on Probability". *The Cambridge Magazine* 1(1): 3-5.
[12] Ramsey, Frank, (1926). "Truth and Probability". Published posthumously in Ramsey, 1931, op. cit.
[13] Ramsey, Frank, (1927). "Facts and Propositions". *Aristotelian Society Supplementary Volume* 7: 153-70.
[14] Putnam, Hilary, (1992). "The nature of mental states". In *The philosophy of mind: Classical problems/contemporary issues*, The MIT Press, 1992, 51-58.
[15] Millikan, Ruth Garrett, (1989). "Biosemantics". *The journal of philosophy* 86(6): 281-297.
[16] Horwich, Paul, (1998). *Meaning*. Oxford University Press.
[17] O'Leary-Hawthorne, John, & Price, Huw, (1996). "How to stand up for non-cognitivists". *Australasian Journal of Philosophy* 74(2): 275-292.
[18] From some notes Ramsey made in his undergraduate times.
[19] Fichte, Johann Gottlieb, (1982 [1797]). "First Introduction to the Science of Knowledge". In *The Science of Knowledge*, Cambridge University Press, 1982, 3-28.
[20] Keynes, John Maynard, (1972 [1930]). "F.P. Ramsey". *The Economic Journal* 42(172): 140-57, page 154.
[21] From his lectures mentioned in note ix.
[22] From his lectures mentioned in note ix.
[23] Misak, Cheryl, (2013). *The American Pragmatists*. Oxford University Press.
[24] Misak, Cheryl, (2018). *Cambridge Pragmatism: From Peirce and James to Ramsey and Wittgenstein*. Oxford University Press.
[25] Schumpeter, Joseph, (1933). "Review of Keynes' *Essays in Biography*". *The Economic Journal* 43(172): 652-7.

EDITORS

EXTERNAL REVIEWERS