Undergraduate Philosophy Journal *of* Australasia

UPJA

Volume 5, February 2024

Founded in 2019, the *Undergraduate Philosophy Journal of Australasia* (UPJA) is the first undergraduate philosophy journal run by students from Australasia. We publish one volume and host two conferences annually and interview philosophers with a substantial connection to Australasia. We aim to be an inclusive and diverse journal and welcome submissions from undergraduates (and recent graduates) worldwide, on any philosophical topic, so long as the author attempts to make a substantive contribution to contemporary philosophy. Submissions from women and other members of underrepresented groups in philosophy, including those for whom English is not their first language, are particularly encouraged.

Undergraduate Philosophy Journal of Australasia – UPJA

@ UPJA_Philosophy

✉ upja.philosophy@gmail.com

www.upja.online

ISSN: 2643-3146

Editors

Brigitte Assi, Australian National University

Brigitte is a current MPhil student at the Australian National University. Their interests are in sexuality, gender, psychoanalysis, phenomenology, and genealogy. Brigitte's Master's project aims to synthesise areas of gendered genealogy, embodied sexuality, and psychoanalysis, to give a distinct account of sexuality and its subversive potential. Primarily looking at Lacan, Butler, Foucault, Irigaray, and Santner, Brigitte situates sexuality as a distinctly subversive feature of the subject, due to its constitutive role.

Billie Angus, Victoria University of Wellington

Billie is a Masters student in Philosophy at Te Herenga Waka: Victoria University of Wellington. Their thesis is focusing on Existentialism, Authenticity, and Social Media. Billie's main interests in philosophy include: Metaphilosophy, Continental Philosophy, and Philosophy of Gender.

James Robinson, Australian National University

James is currently an honours student at the Australian National University where he is researching topics related to risk and ethical decision-making under uncertainty. More broadly, he is interested in ethics, decision theory, and epistemology. After completing his honours, James intends to study a Master's degree overseas next year before undertaking his PhD.

Jean Hew, Nanyang Technological University

Jean Hew graduated from Nanyang Technological University with a Bachelor of Arts (Hons) in Philosophy. Her philosophical interests include epistemology, aesthetics, Chinese philosophy and gender; she wrote her final year paper on the relationship between state power and sexuality. Jean currently works at Jom, a weekly magazine about Singapore.

Leon (Chenglong) Yin, The University of Sydney

Leon is currently an Honours student at The University of Sydney. His honour thesis is on Plato, Foucault, and *Techne*. Leon's research interests include: 20thC Continental Philosophy, Ancient Greek Philosophy, and Philosophy of Sex.

Referees

Ruby Allen, The University of Queensland

Ava Broinowski, The University of Sydney

Christian Carbonell, The University of Valencia

Chow Zhen Yi, Nanyang Technological University

Edwin Chia-en Chen, Macquarie University

Liam Correll, The University of Sydney

Eleanor Edstrom, The University of Sydney

Sanae Ejjebli, Mohamed I University

David He, The University of Sydney

Hugo Heagren, King's College London

Kida Lin, The University of Oxford

Victor Lusis, The University of Canterbury

Evelyn Maude, The University of Toronto

Tessa McKenna, Monash University

Kai Milanovich, The University of Missouri-Kansas

Grier Rollinson, Te Herenga Waka: Victoria University of Wellington

Lindsey Roberts, Macquarie University

William Smith, Australian National University

Editors' Note

2023 for UPJA has been a significant year. We started off the year with a team of four, Eloise Hickey, Mark Rothery, Brigitte Assi, and Rashna Farrukh. We now finish the year with a team of five, with Brigitte Assi assuming the role of Editor-in-Chief, and four other members coming on board. They are Billie Angus, James Robinson, Jean Hew, and Leon Yin. We are incredibly proud of this team, with members coming from Australia, New Zealand, and Singapore. We have all put in a tremendous effort to release the biggest journal so far produced by UPJA, with nine impressive essays in a range of topics.

UPJA prides itself on being an inclusive journal, open to all students in philosophy to meaningfully participate in its creation, whether that be in an Editor role, an author, a referee, or an attendant in our conferences. We are very grateful for all students who enthusiastically engage in the work we do at UPJA, and our journal could not operate without them. We are extremely proud of the expansive range of topics we have included in our fifth journal. The showcase of many different philosophical schools of thought highlight the diverse philosophical interests held by students across Australasia. We are proud to announce that the winner of Best Paper goes to Beau Kent for their paper *"I'm the same – but I'm not": Transracial Adoptees, Hermeneutic Injustice, and Coalitional Politics.* The winner of Best Paper (Member of an Underrepresented Group in Philosophy) goes to Jemma Cusumano for their paper *Diotima's Laughter: Towards a Philosophical Approach Which Centres Ethics*. Both papers argued original philosophical insights eloquently and succinctly, and we are proud to publish them in our fifth volume.

In 2023 UPJA hosted our first ever hybrid conference at the Australian National University. It was very exciting to host students who presented their work in person, and students across the country and globe who presented their work online. When the opportunity arises, UPJA hopes to continue hosting hybrid conferences. UPJA is also launching an audio interview series, to supplement the great interviews already done in our Conversations from the Region series. We hope to continue in our mission to make philosophy more accessible and enjoyable for all those who wish to engage with it. With that, we are looking forward to 2024, and hope you are all excited for what is to come in the future.

Table of Contents

* Winner of Best Paper

† Winner of Best Paper (Member of an Underrepresented Group in Philosophy)

Undergraduate Philosophy Journal of Australasia www.upja.online
VOL. 5, FEBRUARY 2024

Ned Block's homunculi-headed robot and functionalism

JACK BLACKMAN[1]
VICTORIA UNIVERSITY OF WELLINGTON

Abstract

Ned Block posed his thought experiment of the homunculi-headed robot in his paper *'Troubles with Functionalism'* to try to defeat functionalism, a leading theory within the philosophy of mind, which concerns the nature of mental states. The robot was meant to defeat functionalism by showing how functionalism attributes mental states inappropriately, as beings such as the robot would have had mental states under functionalism, despite possessing no qualia. Block's argument rests upon two incorrect assumptions of qualia that this paper exposes as being incorrect. Firstly, Block presupposes that due to qualia being an innate part of human beings, it cannot be accounted for under functionalism. Secondly, Block applies too narrow a perspective onto what qualia can be, and where and how it can subsist, to be considered valid. I attack these two incorrect presuppositions to exhibit how Block's robot cannot be considered to have defeated functionalism.

1. Introduction

Ned Block's homunculi-headed robot is a thought experiment that was posed to exhibit how functionalism is unreliable in posing a paradigm that explains mental states and their relationship to consciousness. The purpose of Block's robot was to

[1] Jack is a student at Te Herenga Waka, VIctoria University of Wellington, studying towards a law degree, and a bachelor of arts degree, majoring in philosophy. He is primarily interested in philosophy of mind, AI, and phenomenology.

highlight how functionalism was too liberal in attributing mental states to beings. Block's robot has the same physical appearance as a human, and similarly exhibits the same behaviours as a human. However, it replaces the human brain with a multitude of little people who are tasked with executing certain operations given certain inputs, which mechanically compels the robot to behave in certain ways. This thought experiment was designed to embarrass the functionalist account of mental states, exposing how they do not accurately explain how mental states engender consciousness. This paper disagrees with the conclusion Block's robot comes to, due to a perceived mistake in how Block perceives the nature of qualia, and his imaginative restrictions on what qualia can be. To suffice this thesis, this paper is split into three distinct sections, which will be provided below, for the sake of clarity:

1. Functionalism: What it is, how we can come to understand it through real life accounts, and how Block's robot relates to it.
2. The innate nature of qualia: What qualia is, what it means for it to be innate, how this relates to the theory of functionalism, and a thought experiment to exhibit how functionalism can possess innate components.
3. Imaginative restrictions on qualia: How our own conscious bias affects our construction on qualia, the wiring of our brains, and the mistake made in common inferences of consciousness.

By the conclusion of this paper, it should be evident that Block's homunculi-headed robot does not defeat functionalism.

2. Functionalism

Functionalism is a well-established theory within the discipline of philosophy of mind, which focuses on the nature of mental states. It is functionalism that Ned Block attempted to confute with his robot. An example of a mental state is possessing a belief, having a desire, being in pain, etc. Functionalism asserts that "what makes something a mental state of a particular type does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part".[2] Functionalism is like a tap; a tap is a tap regardless of what it is made out of, so long as it performs its proper function of dispensing water. Hence, mental states are "fundamentally relational", characterised by relationships with things such as "stimuli, output behaviours, and other mental states".[3] Before the paper proceeds, it is vital that three concepts; consciousness, force, and states, are defined, as they are incredibly relevant as the paper progresses. These will be provided below:

> **Consciousness.** The qualitative experience a being possesses, granted they have a cognitive system consisting of distinctive states which act on and influence each other.
>
> **Force.** The essence of a particular state.
>
> **States.** The realisation of particular conditions which serve independent functions yet are also influenced by other states acting within the same cognitive system of consciousness.

Furthermore, this paper will assume the truth of functionalism, and its tenets as pertinent to this essay. As Block set out to destroy functionalism, and this paper critiques Block, the sake of simplicity demands that this paper assume the truth of functionalist theories.

2.1 Pain and functionalism

To root the theory of functionalism in reality, imagine you were experiencing the sensation of pain. Under functionalism, it is not enough for a being to simply experience the sensation of pain, for such a state to be considered a state. It must experience the state of pain through a system of causal relations, by which the experience of pain is but a constituent within it. Therefore, pain can be characterised simultaneously with other internal states, such as nerves, behavioural outputs, and emotional responses to physical reactions. All of this takes place within a causal network, where the aforementioned internal states are linked to each other and the

[2] Levin, Janet (2004) 'Functionalism', *Stanford Encyclopedia of Philosophy*.

[32] Kobes, Bernard (2007) 'Functionalist Theories of Consciousness' in *The Oxford Companion to Consciousness.* Oxford.

inputs (damage to the body) and outputs (avoidance behaviours), are abstracted from its material realisation.

2.2 Block's homunculi-headed robot

Philosopher Ned Block's paper *Troubles with Functionalism* is often cited as one of the most influential objections towards functionalism.[4] Block proposed the homunculi-headed robot thought experiment as a *reductio ad absurdum* argument to exhibit how functionalism unrealistically accounts for mental states, through an absence of qualia. Block implores the reader to imagine a figure externally analogous to a human being, yet internally differing vastly in nature from humans due to a radically dissimilar constitution (this figure will from hereon be referred to as the robot).[5] Rather than possessing organs and a nervous system, this body contains a multitude of little people, each tasked with executing a specific action given a specific input.[6]

2.2.1 Interacting with the robot.

Suppose that you encountered this robot and asked it how their day was. The robot smiles, telling you that it had a good day, and then begins to inquire into how your day was. This interaction by all societal conventions will be considered perfectly orthodox, yet what's truly absurd about this interaction is that the robot was not acting with an independent whim. Instead, the robot had a society of little people pressing specific buttons which led the robot to perform specific activities. These buttons were illuminated when certain external cues were prompted.

2.3 The robot, and functionalism

In this thought experiment, the robot can perform a task just as any normal person can. However, under the functionalist doctrine, the robot will be said to have mental states, as the states partake in a role within an internal constitution. When analysing the composition of these supposed mental states, Block's robot is meant to expose

[4] Block, Ned (1978) 'Troubles with Functionalism', *Perception and cognition issues in the foundations of psychology*, **9**: 261-325.

[5] Block, 278.

[6] Block, 278.

the absurd conclusions that functionalism entails. Block stresses that there is no "independent reason to believe in the mentality of the homunculi-head".[7] If the robot cannot feel pain, or experience qualia, then the assertion that "the nature of qualia is to be found in its functional role is erroneous, and functionalism is false".[8]

2.4 Functionalism and consciousness

This paper asserts that Block's argument is that functionalism is guilty of liberalism in assigning mental states, and hence, consciousness, to beings. If Block's argument is true, functionalism permits situations where there is "functional equivalence without qualitative equivalence", meaning that qualia "escapes functional explanation".[9] This collapses the structure of functionalism, by refuting the theory that mental states can be understood by virtue of their collective organisation, rather than their internal constitution. This paper takes a leap forward from here, suggesting that Block purports to restrict the scope of consciousness, by advocating for an understanding of qualia as an innate force. If qualia is an innate force, it cannot be accounted for under functionalism, meaning functionalism cannot be true. This paper will defeat this view.

3. Qualia

This section of the essay will explain qualia in detail and exhibit how it plays a fundamental role in the tension between Block's robot and functionalism. Qualia (quale singular) is a term used to denote states of subjective, conscious experience. At any given point, we are experiencing a multitude of sensations; we see letters on the screen, feel our body pressed down against a chair, perceive the colour yellow, feel the emotion of happiness. In each of these cases, we are "the subject of a mental state with a very distinctive subjective character".[10] There is some special

[7] Block, 456.

[8] Walsh, Jackie. (2017) 'Can a Functionalist Account for Qualia', *Oxford Philosophy Society.*

[9] Kind, Amy (n.d.) 'Qualia', *Internet Encyclopedia of Philosophy.*

[10] Tye, Michael (2021) 'Qualia', *Stanford Encyclopedia of Philosophy*.

phenomenological character that defines what it is *like* to undergo these aforementioned states. Broadly, qualia pertains to the quality of our human existence that is accessible solely by virtue of introspection, and experience. Qualia is an incredibly important concept to consider germane to the robot. Whilst with little to no difficulty we can imagine the functional operation of the robot, it is incredibly difficult to believe that the robot experiences qualia. Block's objection is established upon the proposition that whilst the functional replication of the robot as a human can be achieved, it lacks the innate qualia that humans possess to qualify as being conscious.

3.1 Qualia as an innate force

As already mentioned, I posit the tension between Block's robot and functionalism to be rooted in the innateness of qualia. But what does this mean? Block's robot seems to presuppose that qualia is a force which can only be bestowed upon someone innately; that is, they are either born with it, or they don't have it. Because of this nebulous nature, no one has been able to exactly explain how, and why, qualia exists, engendering what is known as the 'hard problem of consciousness'. This problem pertains to "how physical processes in the brain give rise to subjective experience".[11] Qualia as a notion describes the character of what it means to exist. It is a vital component of the human experience, and it is something that Block seemingly cannot envision being artificially replicated. As the robot is able to functionally replicate a human, yet it has different internal wiring, Block is unable to attribute qualia to the robot. This position arises out of a belief that the existence of qualia is determined upon the inner wiring of someone/something. As humans possess the necessary wiring to engender qualia (as it is an innate quality of being human), Block has no problem attributing qualia to humans. However, as the robot contains different wiring, despite producing the same output, Block is unable to attribute qualia to the robot.

[11] Berent, Iris (2023) 'The "Hard Problem of Consciousness" Arises from Human Psychology', *Open Mind: Discoveries in Cognitive Science*, **7**: 564-587. https://doi.org/10.1162/opmi_a_00094.

3.2 The innate nature of qualia and functionalism

If we were to not presuppose qualia to be an innate quality of humans, then the robot would not pose any threat to functionalism. Otherwise, the robot's success in its functional replication of a human would logically entail its success in replicating qualia. The general presumption that Block articulates his argument upon is that qualia exists within humans by virtue of us being human. This is incongruent with the theory of functionalism, as under a functionalist model, all states are understood and defined through their causal relations to other states. If qualia cannot be functionally replicated, its existence as a state is not contingent on contemporary states. A state that was purely innate could not be accounted for by functionalism, as it would not be considered a state due to it not having any relation with other states. This argument hence follows that qualia and functionalism cannot co-exist. I shall lay out the linear reasoning in favour of this thought below:

1. Any conscious being must possess qualia,
2. Qualia is an innate force which does not subsist upon contemporary states,
3. Functionalism attempts to ascertain the nature of states by asserting that states can be understood through their causal relations with other states,
4. Qualia cannot be a part of a functionalist model as it is innate, and does not subsist nor rely on contemporary states,
5. Hence, functionalism cannot account for qualia, and therefore cannot explain the subject of consciousness.

3.2.1 Innateness in functionalism.

If it was shown that an innate force could exist within a functionalist framework, then the previous argument would fall apart at the 4th consideration. This paper argues for the standpoint that a state which possesses an innate force can be understood through either a functionalist lens *or* a singular lens. This is achieved by focusing narrowly on the innate force and will be explicated below.

3.3 Instant coffee

This section of the essay will use the example of instant coffee to offer a compelling argument that an innate force can exist within a functionalist paradigm, or a singular model. To exhibit this, consider a teaspoon of instant coffee. Upon observing this coffee, you will notice a physical, tangible substance which can be held and felt. However, if this coffee was to be submerged in boiling water for 10 seconds, then raised out of the boiling water, what would remain instead would be a teaspoon of boiling water infused with caffeine. Coffee possesses the innate disposition to dissolve in boiling water. However, this disposition is not reliant on contingent forces to *exist*. Regardless of the presence of boiling water, the observed grounds of instant coffee will always possess the disposition to dissolve in boiling water. We can infer from this line of reasoning that the possession of an innate disposition is not contingent on relations to external subjects. A subject can possess an innate disposition through mere virtue of existing in isolation to contingent forces, whether they are activated or not. These dispositions can exist idly, or through the specific activation of their quality, i.e., the coffee ground possesses the ability to dissolve in boiling water. Regardless of the *state* of the disposition, the force still exists nonetheless. This line of reasoning is applicable to the theory of functionalism. Particular states can possess innate forces which don't subsist through causal relations but are merely *realised* through causal relations.

3.4 Objections to the coffee objection

There are two possible objections that I can anticipate to the aforesaid coffee objection. These are the analogy as fact fallacy and an argument from the laws of nature. This paragraph will discuss these two anticipated objections separately and discuss why they should be dismissed.

3.4.1 Analogy as fact fallacy

The analogy as fact fallacy could be argued against the instant coffee objection. This is because of the possible perceived mistake in suggesting that qualia and coffee grounds are comparable enough to apply similar principles from each other. As qualia directly concerns first-person, perspectival experience, whereas coffee grounds are simply a substance that undergoes no inferred experience, the principle may seem weak to analogise. In response to this, it is vital to remember that the coffee objection simply proves that something can have an innate disposition that is only realised through relations to other substances, but nonetheless, will exist in the absence of such substances. This principle is directly relevant to qualia existing within a functionalist paradigm. Assuming the presupposition that qualia is innately bestowed upon us, it can still be understood through its relations to contemporary states. The coffee objection does not seek to assert that coffee and qualia are tantamount. What it does assert is that the relationships their dispositions have with other substances can be analogised, to exhibit how an innate disposition can exist both independently, and through other substances.

3.4.2 The laws of nature

Furthermore, there is an argument that the disposition of coffee can be explained through the laws of nature. As it can be explained by something beyond itself, this would mean that it does not have an innate disposition, as it subsists through causal relations, rather than solely being perceived through them. This would contravene the nature of qualia (which is merely realised through other substances) and render the analogy invalid. In doing so, it would affirm Block's argument, showing that qualia are an innate quality that cannot be accounted for under functionalism, hence defeating functionalism. The most logical way to evaluate which objection is stronger; that of the coffee grain, or the argument centred around the laws of nature, seems to be ascertaining whether there is a theoretical arbiter which can identify when something can be considered as innate. This bar will then assert the feasibility of the coffee argument as a valid argument against Block's objection.

3.5 The requisite for innateness

An arbiter of innateness would be a force that can be considered innate within the parameters of the word as imposed in this paper. Once there is an arbiter of innateness, it can be related to qualia, to answer questions about qualia. In finding this arbiter, it intuitively made sense to conduct an inquiry into whether electrons are innate. This is because they are one of the smallest constituents of the universe and are ubiquitous throughout our world. If anything made of matter in our reality was to be considered innate, it would be electrons. Electrons encircle the nucleus of atoms at specific energy levels to influence the character of atoms. Due to this, they are incredibly small in matter, dwarfing that of an atom, meaning it's reliable to utilise as an arbiter for innateness.

3.5.1 The innate quality of an electron

Electrons have rules, dispositions and qualities that constitute what it is to be an electron. Electrons must behave as electrons, as they are in strict obedience to the parameters imposed on them by nature. They underpin everything in the universe, and are an incredibly small force, which cannot be reduced any further beyond themselves. The existence of electrons, and the properties that they occupy, cannot be explained by anything but themselves. They can be perceived and understood through their causal relations to other forces, but they do not subsist upon the activations of these forces for their innate dispositions to exist. For the purposes of this paper, electrons must be considered to be innate. In spite of all this, electrons can be accounted for under functionalist forces.

3.5.2 Qualia and electrons as innate forces

Qualia existing within conscious beings, the notion presupposed by Block which is pivotal to his functionalist attack, is tantamount to electrons existing within the functionalist order of the universe. As electrons and coffee grains can be functionally accounted for within the parameters imposed in this paper, qualia must be treated similarly, ultimately evincing how the innate force of qualia can be accounted for within a functionalist framework.

4. Narrow construction of qualia

I asserted earlier in this paper that Block purports to restrict the scope of qualia, hindering it to a specific organisation which Block appears to maintain as necessary for having mental states, i.e., consciousness. The robot's supposed lack of qualia stems from it not having a brain. It looks the same as us, acts the same as us, and operates the same as us, yet has a queer internal constitution which seemingly prevents it from possessing consciousness. Directly or indirectly, Block asserts that the acquisition of qualia is restricted to those entities with a brain. This discriminates against any other being which could exemplify all the modalities that would lead us to register that being as conscious yet disabling it from being regarded as conscious due to it possessing different internal wiring.

4.1 The necessary wiring

Block's philosophy can be seen to maintain that humans must possess the correct constitution to elicit consciousness. However, when we detach our constitution from how normalised we have learned it to be, it appears to be just as absurd as the robots. Consider this excerpt explaining the present moment from MIT's school of engineering -

> When you read these words, for example, the photons associated with the patterns of the letters hit your retina, and their energy triggers an electrical signal in the light-detecting cells there. That electrical signal propagates like a wave along the long threads called axons that are part of the connections between neurons. When the signal reaches the end of an axon, it causes the release of chemical neurotransmitters into the synapse, a chemical junction between the axon tip and target neurons. A target neuron responds with its own electrical signal, which, in turn, spreads to other neurons. Within a few hundred milliseconds, the signal has spread to billions of neurons in several dozen interconnected areas of your brain and you have perceived these words.[12]

Our consciousness is derivative from processes completely tantamount to the homunculi-headed robot, spare for the instruments utilised to necessitate the functions. We are walking homunculi-headed robots assigned with consciousness, however we cannot even be sure in stating that we really do possess consciousness. While I won't stray from the motive of the paper, it's worth noting that a myriad of

[12] Dougherty, Elizabeth (2011) *What are thoughts made of?* MIT School of Engineering.

theories pertaining to this, such as the simulation theory, solipsism, and Pyrrhonian scepticism, carry valid weight and too trouble Block's theory.

4.2 Reasons to believe in qualia

In reference to his scepticism pertaining to the homunculi-headed robot, Block states that "there is no independent reason to believe in the mentality of the homunculi-head, and I know of no way of explaining away the absurdity of the conclusion that it has mentality"[13]. This must be taken in consideration with Block's acceptance of humans possessing consciousness, as due to us having brains, there is an independent reason to accept our consciousness.

4.2.1 Does everyone have brains?

It is commonly assumed that we all have brains and are conscious, but either of these assumptions could be wrong. This is because attributing consciousness is ultimately provisional without access to the other person's brain. Consider Brian Keeley's distillation of this situation -

> I happen to think, say, that Ted Cruz would make a great US President and I support his candidacy. Of course, one must be a native-born US Citizen to be eligible for that office. I've never seen his birth certificate. So, I continue to believe he would be a great POTUS, but if somebody provided me with evidence that he was, in fact, born in Scotland, then I'd of course revise that opinion. But part of my reason for thinking he'd be a great President is that I think he's U.S. born. But that reason is defeasible. In the same way, my belief that you have a brain is defeasible.[14]

4.3 Phenomenological traits of consciousness

In troubles with functionalism, Block references Nagel, who states that conscious experience exists "if and only if there is something that it is like to *be* that

[13] Block, 456.

[14] Bowen, Connor (2019) *Global Consciousness: A Functionalist Neurophilosophical Perspective*. Claremont McKenna College.

organism"[15]. As much as we can try to ascertain what beings besides humans have consciousness, it is ultimately futile as no amount of physical inference will enable us to conceive what foreign phenomena may be like. While we can postulate that a pigeon is a conscious being, there is no way that we can ultimately be sure of it. Hence, physical constitution can only take you so far in determining a being's consciousness, due to the lack of diverse qualitative experience that we can recount. Due to this, we do not have the sufficient tools to ascertain what it's like to be a homunculi-headed robot. Block contradicts himself in referencing Nagel's philosophy as he cannot provide an accurate description of what it is like to be that robot, vic. he cannot determine whether or not it has qualia. This suspends Block's case in a philosophical impasse. His inability to assert with certainty as to whether or not there is something that it is like for the robot to be a robot, leaves his case unconditionally incapable of proving itself correct.

4.4 Imaginative restrictions on qualia

This is because our conception of consciousness is limited to what we are capable of imagining within the scope of our own consciousness. Because of this, we are fundamentally limited in having a broad perception of what different mental states would be like, and what different conditions of consciousness these mental states could engender. While we can postulate, for example, that inanimate objects are incapable of experiencing qualia, there is nothing we can raise to support this affirmation beyond our blind intuition. Our inability to completely understand our own consciousness, our own qualia, and our own mental states, extends to our inability to epistemically and metaphysically access the existence of these aforementioned states in beings and forms different to us. This bias is why most people, including Block, attribute with certainty that those like us are conscious. Those, like us, who are able to walk, talk, think, express emotion, passion, interest etc. We assume that once organisational prerequisites for consciousness are met, then consciousness shall arise. However, this organisational prerequisite shouldn't be

[15] Nagel, Thomas (1974) 'What is it like to be a bat?', *The Philosophical Review*, **83**(4): 436.

hindered solely to the physical organisation of humans, as we simply cannot purport to maintain consciousness as a phenomena solely experienced by humans.

4.5 Block's restricted imagination

This prior paragraph is essential to understanding the error underpinning Block's philosophy on the robot. To him, it is absurd that the robot be considered conscious, as it is merely a mechanical replication of a human, and therefore the consciousness that accompanies human experience cannot be replicated. Block is unable to imagine that qualia can exist in a way that differs from human qualia.

4.5.1 Human qualia compared to the robots

This is because human qualia is understood as the norm. When you make a witty remark and somebody laughs, you assume to understand the qualitative experience those around you went through when they laughed. However, this assumption is rooted within a fallacy that all humans experience qualia the same. As we only have qualitative, first person access to anything that we personally go through, we are involuntarily suspended in an unfortunate state of incessant scepticism, where we can only infer that others have similar qualitative experience as us, with no robust means of making sure of this.

4.5.2 A joke in a room full of people

To exhibit this in reality, suppose that you made a joke in a room filled with three strangers, and they all laughed. Your assumption of their qualitative experience will be that, like you, they found what you said humorous, felt an impulse to express this physically to you, leading to the appropriate behavioural output to reflect that. However, you later find out that in that room, one of them had schizophrenia, one of them was high on LSD, and the other person experienced inverted qualia. Your assumption of what they felt, thought, and perceived would be terribly inaccurate, as you have no means to assert what their qualitative experience truly was. Simply because someone has a brain does not assert that their conscious experience is analogous to yours. Everybody that we encounter in life and regard as conscious, we do so because they look like us. Without knowing if they are real, if they have a brain, or if they experience life in a sense similar to us, we maintain them as conscious, due to the similarities we share in physiognomy and behaviour. If consciousness could be confidently inferred merely by virtue of having a brain, as Block supposes it to be, all the idiosyncrasies and syndromes of the brain would reflect replicative experiences of life. Block's ideology would anoint consciousness to someone in a coma. Block's thought suffers from a lack of imagination of what it is like for the robot to be a robot. If we treated the organisational setup of a human organisational as Block does the homunculi-headed robot, we'd observe the infinite differences in the organisation of our brain to anyone else's brain. We'd see the utter lack of understanding of how our brain, mental states, qualia, and consciousness works. It would look as absurd to call us conscious, as it is to call the homunculi-headed robot conscious. This imagination we project onto assuming we are conscious, can be extended for us to conceive the homunculi-headed robot conscious.

5. Conclusion

This paper explored Block's homunculi-headed robot and how it ultimately fails as a rejection for functionalism. The subject of qualia was explored in depth to suffice this thesis. Block's mistake in presupposing that due to qualia being innate it is unable to exist within a functionalist system was exposed and proven wrong. Furthermore, phenomenological exploration of the subject of qualia and how our own experiences trap us in a perspective bias buttressed the fact that Block is unable to prove that the robot does not have qualia and hence cannot object to functionalism. It is with confidence that we can assert that Block's homunculi-headed robot does not defeat functionalism.

References

Berent, Iris (2023) 'The "Hard Problem of Consciousness" Arises from Human Psychology', *Open Mind: Discoveries in Cognitive Science*, **7**: 564-587. https://doi.org/10.1162/opmi_a_00094

Block, Ned (1978) 'Troubles with Functionalism', *Perception and cognition issues in the foundations of psychology*, **9**: 261-325.

Bowen, Connor (2019) *Global Consciousness: A Functionalist Neurophilosophical Perspective*. Claremont McKenna College.

Dougherty, Elizabeth (2011) *What are thoughts made of?* MIT School of Engineering.

Kind, Amy (n.d.) 'Qualia', *Internet Encyclopedia of Philosophy.*

Kobes, Bernard (2007) 'Functionalist Theories of Consciousness' in *The Oxford Companion to Consciousness.* Oxford.

Levin, Janet (2004) 'Functionalism', *Stanford Encyclopedia of Philosophy*.

Tye, Michael (2021) 'Qualia', *Stanford Encyclopedia of Philosophy*.

Nagel, Thomas (1974) 'What is it like to be a bat?', *The Philosophical Review*, **83**(4): 436.

Walsh, Jackie. (2017) 'Can a Functionalist Account for Qualia', *Oxford Philosophy Society.*

Diotima's Laughter: Towards a Philosophical Approach Which Centres Ethics

JEMMA CUSUMANO[16]

UNIVERSITY OF QUEENSLAND

Abstract

Plato, a seminal figure in Western philosophy, employed the dialogical method in his writing to underscore the significance of dialectical reasoning and open discourse. In Plato's Symposium, there is an exchange between Socrates and Diotima whereby the latter teaches the former the art of love. The majority of philosophical discussions concerning the exchange typically interpret Diotima's teachings as representative of Platonism and acknowledge the presence of Plato's Theory of Forms within it. However, in Luce Irigaray's analysis of this dialogue, she emphasises Diotima's unique position within the Symposium. Irigaray, in directing her attention to Diotima herself, is able to provide a reading which pays attention to the nuanced moments where Diotima's views transcend the bounds of Platonism. With this reading as my starting point, I argue that Diotima's laughter in her speech promotes an ethical approach to philosophy as a way of life. Paired with her pedagogical approach, Diotima fosters an ethical exchange with Socrates which

[16] Jemma Cusumano is a recent graduate from the University of Queensland, with Honours in Philosophy. Her research interests include feminist philosophy, psychoanalytic philosophy, and ethical theory. Her most recent projects have focused on Hannah Arendt's concept of the banality of evil, and Simone de Beauvoir's *The Ethics of Ambiguity*.

challenges conventional hierarchical and oppositional thought within philosophy. By highlighting Diotima's laughter, pauses, and questioning, Irigaray's interpretation illustrates a philosophical approach which is open to otherness and embodies plurality. In sum, this paper showcases how laughter in Irigaray's reading of Diotima's speech advocates for an ethical foundation in philosophy, emphasising the transformative power of dialogue and the importance of embracing diverse perspectives. It underscores the enduring relevance of Plato's dialogues in inspiring ethical engagement in philosophical inquiry.

1. Introduction

Hailed as one of the founders of Western philosophy, Plato bears immense significance due to his foundational contributions which continue to shape the field today. Notably, Plato's use of the dialogical method in his philosophical works emphasises the importance of dialectical reasoning and open discourse in philosophy. In this paper, I will examine the dynamic exchange that takes place between Socrates and Diotima in Plato's *Symposium*.[17] More specifically, I will focus on the role of laughter in Luce Irigaray's reading of Diotima's speech in her chapter titled 'Sorcerer Love: A Reading of Plato, *Symposium*, 'Diotima's Speech',' from her book, *An Ethics of Sexual Difference*.[18] Firstly, I will outline the importance of Irigaray's reading of Plato and the unique position of Diotima's speech among the male speeches in the *Symposium*. Subsequently, I will present my argument that Irigaray's analysis reveals how laughter within Diotima's speech instils an approach to philosophy as a way of life rooted in ethics. To reach this conclusion, I will first explore Irigaray's emphasis of Diotima's pedagogical approach, which effectively establishes an ethical exchange between her and Socrates, challenging conventional hierarchical and oppositional thought. Subsequently, I will argue that Irigaray's interpretation illustrates how Diotima's laughter creates a momentary pause. This interval in the conversation disrupts Socrates' established truths, fosters an openness

[17] Plato (1967) The Symposium, Walter Hamilton, trans, Penguin.

[18] Irigaray, Luce (1993) An Ethics of Sexual Difference, Carolyn Burke and Gillian C Gill, trans, Cornell University Press.

to others, and reflects the essence of plurality. Ultimately, the role of laughter in Irigaray's reading of Diotima's speech advocates for an approach to philosophy which is founded in ethics.

2. Contextualising Diotima's Speech: Plato's *Symposium*, Irigaray's Reading, and the Philosophical Landscape

Before I discuss the role of laughter in Irigaray's reading of Diotima's speech, I will briefly explain Plato's *Symposium* and the significance of Diotima's speech within it. Plato's *Symposium*, like numerous other works of his, was written in the form of a dialogue, whereby the characters partake in a give-and-take interchange.[19] This work depicts a gathering of men at a banquet in ancient Greece, who are prompted by Eryximachus to present, one by one, an encomium—a speech that praises Love (Eros).[20] In this way it is unlike many of Plato's other works as it is made up of a series of speeches from different characters who either comply with the challenge, or take different approaches to the topic. This diversity allows readers to consider various philosophical perspectives on several key philosophical ideas, such as Plato's theory of forms and the ladder of love. It also delves into the connection between love and beauty, as well as the pursuit of higher knowledge. The *Symposium* is considered by many as the source of many Western interpretations and analyses of love. Diotima's speech within the *Symposium* is unique as she is the only female character to be given a voice amid the male speeches. Nevertheless, there is considerable scholarly debate as to whether Diotima's voice can genuinely be considered hers, given that it is conveyed through Socrates, who describes Diotima as the prophetess who taught him the art of love.[21] In contrast to these analyses, which often scrutinise Diotima's gender and question the fidelity of Plato's representation, Irigaray takes a different approach. She deliberately avoids attributing the speech to either Plato or Socrates and instead interprets it as the authentic expression of Diotima. Tina Chanter, in her work *Ethics of Eros: Irigaray's*

[19] While there is open debate on why Plato chose to write the Symposium in a dialogical form rather than a single speech, I take it as a given that through this dialogue form, Plato's evident interest in pedagogical questions is demonstrated.

[20] Plato, The Symposium, 40–41.

[21] Plato, The Symposium, 79. This debate mirrors historical gender dynamics within the philosophy profession, which has been predominantly male until relatively recently.

Rewriting of the Philosophers, interprets this approach as a means of returning Diotima's agency and emphasising 'the uncertainty that surrounds not only Diotima's words, but her very existence.'[22] By presenting Diotima's words as her own, Irigaray not only challenges traditional interpretations but also underscores the broader issue of the exclusion of women in philosophical discourse. This aligns with the overarching goal of Irigaray's work, emphasising the ongoing uncertainty and the need to secure a place for women's voices within the philosophical canon.

For most interpretations of Diotima's speech in Plato's *Symposium*, the focus is usually on Diotima's delineation of the so-called ladder of love, where knowledge is the final destination of the ascent toward Beauty.[23] However, because of the central task of her book, *An Ethics of Sexual Difference*, Irigaray's reading of Diotima's speech is deliberately subversive. Her project involves revisiting texts from the philosophical canon to re-evaluate how they have been interpreted in order so 'we might begin to rethink human being in terms of' relations rather than oppositions. For Irigaray, this rethinking serves as the foundation for her theories of sexuate difference, and as Rachel Jones argues, 'would necessarily transform philosophy, re-orienting our approach to fundamental philosophical questions about the origin of being, and the relation of form and matter.'[24] Because of Irigaray's method of rereading which involves looking for nuance and ambiguity within these texts, Irigaray locates Diotima's contribution to the overarching theme of love's diverse manifestations and philosophical significance in an earlier forgotten passage. Irigaray reads Diotima as arguing for the intermediary nature of love. This reading goes against the metaphysical trajectory of which the *Symposium* seems to support, as Irigaray's Diotima teaches of a logic of relation rather than one of opposition.[25] This aligns with the overarching project of Irigaray's book which argues that the two

[22] Chanter, Tina (2016) Ethics of Eros: Irigaray's Rewriting of the Philosophers, Routledge, 162.

[23] Jones, Rachel (2011) *Irigaray*, Polity Press, ProQuest Ebook Central, 80. In the Stanford Encyclopedia article titled 'Plato on Friendship and Eros,' C. D. C Reeve states that 'what [Diotima] teaches [Socrates], in a nutshell, is Platonism.' This reflects the prevailing view that Diotima is perceived as a conduit through which Plato articulates and presents his theory of Platonic Forms: Reeve, C D C (2023) 'Plato on Friendship and Eros' in Edward N Zalta and Uri Nodelman, eds, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.

[24] Jones, *Irigaray*, 83.

[25] See Rachel Jones' discussion in her book *Irigaray* for a more in-depth discussion of love's role as intermediary as interpreted by Irigaray.

different sexes are not two copies or versions of the same, but relational beings who exist as two.

Importantly, Irigaray's reading of Plato is a subtle one. She does not claim that it is absolutely true or correct for that would go against her project which denies the idea that there is a single truth that is one and the same for all. In her book *Slow Philosophy*, Michelle Boulous Walker discusses the significance of Irigaray's reading in her chapter titled 'Rereading: Irigaray on Love and Wonder.' Boulous Walker claims that Irigaray skilfully avoids homogenising Diotima's message, instead highlighting the ambiguities and tensions that constitute its complex otherness.[26] Boulous Walker astutely recognises Irigaray's approach to philosophy as an ever-evolving, dynamic process, one that eschews rigidity and completeness in favour of perpetual transformation.[27] I think this is exemplified by Irigaray's deliberate avoidance of a position of critique. While she acknowledges shortcomings in Diotima's method at various points,[28] Irigaray does not judge this as a sign of fault. Instead, she interprets these shortcomings as indicative of the ambiguous and plural nature of Diotima's voice and message. Although I do not have the scope to explore the broader implications of Irigaray's theory, this contextual background is essential for comprehending the role of laughter in her interpretation of Diotima's speech.

3. Laughter in Philosophy

So, how does Irigaray interpret the laughter in Diotima's speech? While her discussion of it is brief, it is crucial to understand that Irigaray views Diotima's laughter directed at Socrates as an interaction devoid of hostility or anger.[29] According to Irigaray, Diotima's laughter is not a reprimand but rather a gentle chiding aimed at Socrates for his misunderstanding of the intermediary nature of

[26] Boulous Walker, Michelle (2016) *Slow Philosophy: Reading against the Institution*, Bloomsbury Publishing, 78.

[27] Boulous Walker, *Slow Philosophy*, 79.

[28] Irigaray, An Ethics of Sexual Difference, 27, 29.

[29] Irigaray, An Ethics of Sexual Difference, 22.

love.[30] She laughs at his mistaken assumption that because 'everybody admits that he is a great God,'[31] love cannot be ugly or bad. This reading of Diotima's laughter is significant because it stands in contrast to conventional readings of laughter in Plato's work. In her article titled 'The Laughter of Hannah Arendt,' Boulous Walker argues that many scholars believe ancient Greeks took laughter seriously, often associating it with ridicule and a humiliating lack of respect.[32] Plato consistently expressed his disapproval of laughter and humour, considering it an emotion that overrode rational restraint and was tinged with malice. In his work *Philebus*, he scrutinises comedy as a form of mockery: 'In laughing at them, we take delight in something evil—their self-ignorance—and that malice is morally objectionable.'[33] This assessment of laughter aligns with John Morreall's claim that all laughter in Plato's work is aimed at self-ignorance.[34] This corresponds to the superiority theory of laughter, which posits that laughter expresses a sense of superiority either over others or over our previous selves.[35]

Nonetheless, Boulous Walker challenges the idea that the laugh of ridicule is the only form of laughter found in Plato's works.[36] Drawing from Arendt's interpretation of Plato, she views the laughter of the Thracian maid as an example of innocent laughter, a manifestation of common sense.[37] Similarly, in her analysis of Irigaray's reading, Boulous Walker characterises Diotima's laughter as playful mockery, a mode of interaction that, in her view, distinguishes itself from the more confrontational exchanges among the male participants at Plato's *Symposium*.[38] This characterisation of Diotima's laughter as a light-hearted teasing is justified by

[30] Plato, *The Symposium*, 80.

[31] Plato, *The Symposium*, 80.

[32] Boulous Walker, Michelle (2021) 'The Laughter of Hannah Arendt,' *ABC Religion and Ethics*, https://www.abc.net.au/religion/the-laughter-of-hannah-arendt/13401584.

[33] Plato (1978) *The Collected Dialogues of Plato*, Edith Hamilton and Huntington Cairns, trans, Princeton University Press, 48–50.

[34] Morreall, John (1982) 'A New Theory of Laughter,' Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition **42**, 243, doi:10.1007/bf00374037.

[35] Morreall, 'A New Theory of Laughter,' 244.

[36] This exemplifies a common misconception where people conflate Plato and Platonism, mistakenly considering them synonymous. In reality, Platonism is a reading and interpretation of Plato's work. Irigaray's reading highlights that with careful attention, we can discern the distinctions between them.

[37] Boulous Walker, 'The Laughter of Hannah Arendt.'

[38] Boulous Walker, *Slow Philosophy*, 80.

Diotima's use of various other pedagogical methods. These include questioning and taking pauses to attentively listen to Socrates' responses, aspects I will explore in detail later on. Therefore, Diotima's playful teasing stands in stark contrast to the laughter of superiority, which may be hostile and ridiculing. By emphasising the other pedagogical methods Diotima employs in her discussion with Socrates, Irigaray is able to put forward an interpretation of Diotima's laughter which departs from the norm, which is the other more aggressive types of laughter often found in Plato's works. This underscores the unique significance of laughter in Diotima's speech. Comprehending this contrast allows for a more profound exploration of the subtleties in Irigaray's reading in relation to established philosophical traditions.

4. Challenging Conventional Hierarchical and Oppositional Thought

Using Irigaray's interpretation, I will now illustrate how Diotima's laughter lays the foundation for an ethical framework, nurturing an approach to philosophy rooted in ethics. I see the ethical dimension of Diotima and Socrates' exchange as primarily established through Irigaray's emphasis on Diotima's pedagogical approach, which centres around questioning. Irigaray's Diotima signifies this pedagogical method by pairing her gentle laughter with an open disposition, which involves raising questions for Socrates to answer rather than dictating what he should know. On Irigaray's account, Diotima's laughter should not be seen as undermining an open discourse, for it is not angry, but rather, she laughs to dismantle Socrates' assurance of opposing terms.[39] This is exemplified when Diotima asks Socrates what he thinks the nature of love is. By affording Socrates the chance to respond and actively listening to what he has to say, Diotima grants him a voice in the discussion.[40]

[39] Irigaray, *An Ethics of Sexual Difference*, 22. This marks the fundamental divergence between Diotima's approach to questioning and that of Socrates. While both engage in questioning for philosophical exploration, their approaches differ in terms of context, subject matter, educational focus, and style. Here, Diotima's unique use of laughter, as described by Irigaray, suggests a method of deconstruction, challenging the binary thinking inherent in Socratic dialogues. This stands in contrast to Socrates, who predominantly employs questioning as a means of uncovering truth and fostering understanding. Here, the intriguing reversal occurs, as Socrates, once the educator, becomes the subject of education through Diotima's insightful questioning.

[40] I acknowledge that there is potential for Socrates to have misinterpreted Diotima's intent in laughing at him. However, I would argue that Diotima prevents this from happening by continually prompting Socrates to respond to her questions, making it clear that her laughter is not at all angry but a tool for dismantling his assurances.

Irigaray describes this dynamic as a 'dialogical volleying between Diotima and Socrates.'[41] This process of inquiry prompts Socrates to reconsider his earlier conviction that love is a god, ultimately setting the stage for Diotima to introduce her argument for the demonic nature of love.[42] For Irigaray, it is not because love lacks the beautiful and the good things that he loses his status as a God, but because love is an intermediary, 'neither mortal nor immoral,' but a state between them both.[43] By guiding Socrates to this conclusion through questioning rather than assertion, Diotima avoids assuming a position of mastery and establishes herself as an equal participant in collaborative inquiry. While Diotima does steer the conversation, her use of laughter underscores her amiable demeanour, mitigating any sense of superiority over Socrates. This pedagogical approach, which Irigaray calls attention to, promotes an open dialogue between Diotima and Socrates, fostering a mutually respectful exchange of ideas where neither holds power over the other.

To elaborate on how Diotima's pedagogical approach dismantles hierarchies, I find resonance in Hannah Arendt's perspective in *The Life of the Mind*. Arendt views laughter as a force that disrupts the overly rigid distinctions between the 'common man' and the 'speculative thinker,' blurring the boundaries between the many and the few.[44] In Boulous Walker's analysis of Arendt's work, she interprets Arendt's use of laughter as a manifestation of common sense's response to philosophical thought, acting as a reminder and remedy of the limits of excessive rationality.[45] This is evident when Arendt writes: 'Laughter rather than hostility is the natural reaction of the many to the philosopher's preoccupation and the apparent uselessness of his concerns.'[46] Similarly, Irigaray views Diotima's laughter as a response to Socrates' lack of common sense: 'She continues to laugh at his going to look for his truths beyond the most obvious everyday reality, at his not seeing or even perceiving this reality.'[47]

[41] Irigaray, An Ethics of Sexual Difference, 25.
[42] Plato, *The Symposium*, 81.
[43] Irigaray, An Ethics of Sexual Difference, 22.
[44] Arendt, Hannah (1978) *The Life of the Mind*, Secker & Warburg, 81.
[45] Boulous Walker, 'The Laughter of Hannah Arendt.'
[46] Arendt, The Life of the Mind, 82.
[47] Irigaray, An Ethics of Sexual Difference, 26.

Whilst I recognise the importance of highlighting common-sense experiences when doing philosophy, I find the role of laughter in deconstructing the division between the 'philosopher' and the 'common man' to be particularly relevant. This, I believe, directly challenges the prevailing Western philosophical tendency to favour one side of an apparent duality. This inclination underscores the hierarchical distinction within dualisms such as philosophers and commoners, logic and rhetoric, and teacher and student. While Arendt does not explicitly endorse the notion of philosophers standing on equal footing with the many, her depiction of the 'intramural warfare' between thought and common-sense hints at the possibility of challenging these hierarchies. She argues that the historical practice of philosophy often involves philosophers problematically detaching themselves from the common world for extended periods.[48] In her discussion of why this detachment is problematic, I discern a suggestion that the preference for withdrawal is a manifestation of the tendency to establish hierarchies within philosophy. If Diotima were to conform to this hierarchical tradition, she would inherently possess power and authority over Socrates due to her roles as a teacher and philosopher. Being a symbol of elevated thinking, conventional hierarchical norms would position her as superior to Socrates, who embodies the archetype of the 'common man.' This understanding of Diotima's speech often emerges in traditional readings of Plato, wherein her lesson is construed as detailing the hierarchical ascent of love.

In contrast, Irigaray, through an emphasis on the relational facets of Diotima's pedagogical method—encompassing questioning and laughter—illustrates how these approaches effectively dismantle such hierarchical distinctions. This subversive interpretation by Irigaray suggests that Diotima is instructing Socrates on how love navigates between opposing elements, intertwining them rather than establishing one above the other.[49] Irigaray's emphasis on Diotima positioning herself as Socrates' equal aligns with what I see as Arendt's underlying aim to challenge conventional hierarchical approaches to philosophy. The significance of establishing this form of

[48] Arendt, The Life of the Mind, 81.
[49] Jones, *Irigaray*, 82.

exchange within philosophy is that it is ethical in a relational way.[50] In the works of both Irigaray and Arendt, laughter emerges as a valuable tool for fostering a reciprocal (or even loving) relationship between two individuals engaged in philosophical dialogue. This non-hierarchical exchange treats participants as equals, fostering a more enriched and inclusive discourse, as well as a mutually beneficial learning process that embodies ethical virtues such as equality and inclusivity. The importance of these virtues lies in their potential to cultivate a positive and ethical learning environment, and reflect an approach to philosophy firmly centred on relational ethics.

Diotima's laughter, as interpreted by Irigaray, not only challenges hierarchical thinking and the tendency to favour one side of a duality but also transcends the duality itself. Her pedagogical approach, which blends questioning and laughter, fosters a collaborative relationship between herself and Socrates which defies the adversarial mode of interaction and thinking commonly found in the dominant philosophical tradition.[51] In this tradition, interactions often involve opposition and tension, with one side of the dialogue assumed to be authoritative and correct, and the other considered incorrect. What, in my view, sustains such oppositional thinking is the dominating view of philosophy as primarily a desire for knowledge, rather than a love of wisdom.[52] While I do not have the scope to explain in depth the issue of these different approaches to thought, it is essential to grasp, albeit in broad strokes, how they shape our philosophical methods. Boulous Walker characterises the former approach as 'a philosophical tendency … that stifles ambiguity and

[50] The main reason underscoring why dismantling hierarchies is important is because it challenges the traditional notion that there is a single truth and knowledge that is one and the same for all. This reflects Irigaray's project to re-orient our approach to philosophy and knowledge. While I understand that in certain pedagogical settings, maintaining a clear hierarchy may be advantageous, it is essential to recognise that my claim about laughter dismantling hierarchies within pedagogical settings does not necessarily imply a blanket disregard for hierarchies. My argument primarily focuses on situations where laughter can contribute to a more inclusive and collaborative learning environment. In cases where a clear hierarchy is necessary, such as when addressing academic disparities or managing disruptive behaviour, the application of laughter as a pedagogical tool can be adapted to suit the specific needs of the classroom and foster a conducive atmosphere for learning.

[51] The relational nature of Diotima's laughter fosters a sense of equality and encourages open dialogue as a means of doing philosophy, and is reflected in her focus on 'love' as an intermediary that entwines rather than establishes hierarchies.

[52] The problem with this approach to thought is the central topic of Boulous Walker's book, *Slow Philosophy: Reading against the Institution*.

uncertainty (otherness) beneath layers of knowledge.'[53] This approach, with the dominating principles of system and certainty, prioritises the end result—the conclusion—rather than the process which leads to that conclusion. By doing so, it tends to reduce philosophy to 'a forensic practice of searching out flaws in arguments.'[54] As previously discussed, this confrontational style of discourse is exemplified by the male participants in Plato's *Symposium*. This is exemplified soon after Socrates concludes his speech, as Aristophanes endeavours to argue that Socrates had referenced his theory at a certain juncture.[55] In this instance, we observe a distinct 'desire to know' approach to philosophy, characterised by the competitive assertion of one's perspective and the endeavour to establish intellectual dominance through argumentation. Additionally, in such instances where a desire to know is prioritised, laughter may function as a formidable weapon used to outrightly discredit opposing perspectives.

In contrast, when a love of wisdom is prioritised within philosophical work it defines philosophy as 'a way of life that binds philosophers to philosophy.'[56] Irigaray's reading demonstrates this approach to thought as she finds within Diotima's message models of engaged and ethical encounters rather than an exhaustive and systematic theory of Platonic Forms. According to Irigaray, Diotima's laughter works to dissolve the tension that typically arises from conflicting viewpoints. Instead of engaging in confrontational argumentation with Socrates and getting entangled in his metaphysical grappling, Diotima establishes a collaborative rather than combative dialogue. This approach reflects a love of wisdom not only in the content of the philosophical message but also in the method of engaging with others—fostering a connection between the philosopher and philosophy that goes beyond oppositional debates. In Irigaray's Diotima I see a commitment to the shared pursuit of knowledge and understanding rather than a focus solely on individual perspectives.

[53] Boulous Walker, *Slow Philosophy*, 92.
[54] Boulous Walker, *Slow Philosophy*, 4.
[55] Plato, *The Symposium*, 96.
[56] Boulous Walker, *Slow Philosophy*, 2.

Irigaray introduces this perspective at the beginning of her interpretation, stating: 'Diotima's teaching will be very dialectical, but different from what we usually call dialectical. In effect, it doesn't use opposition to make the first term pass into the second in order to achieve a synthesis of the two.'[57] In other words, Diotima's teaching is unique in that it does not follow the traditional dialectical pattern of setting up two opposing ideas and then reconciling them to reach a conclusion. Her method does not rely on opposition as the driving force for synthesis; instead, it 'unveils the insistence of a third term that is already there and that permits progression.'[58] For Diotima, this intermediary is love. She employs laughter as a tool to disrupt the conventional, oppositional modes of philosophical discourse, creating a more collaborative and open atmosphere. Laughter is not employed to refute or ridicule; rather, it serves as a means for both Diotima and Socrates to gather their thoughts and engage in a more harmonious and mutual exploration of the concept of love. Because her laughter embodies a playful form of mockery, Irigaray views Diotima's teaching method as turning questioning into a joyful and positive experience, rather than something to be feared.[59] This establishes an ethical exchange between her and Socrates because it constitutes a dialectical approach that promotes the sharing of ideas and the cultivation of understanding without the need for rigid opposition. Consequently, in accordance with Irigaray's interpretation, Diotima's laughter serves to depart from the traditional confrontational dialectical approach in philosophy, thereby nurturing an ethical exchange that characterises philosophy as a way of life rooted in ethics.

5. Ethical Inquiry, Unlearning, and Plurality in Diotima's Pedagogy

Diotima's laughter, as read by Irigaray, also embodies a philosophical approach with ethics at its core, as it prompts Socrates to reconsider his deeply entrenched beliefs. Irigaray claims that Diotima 'ceaselessly examines Socrates on his positions but without positing authoritative, already constituted truths.'[60] The function of laughter within this pedagogical method lies in its ability to evoke a sense of bewilderment

[57] Irigaray, An Ethics of Sexual Difference, 20.
[58] Irigaray, An Ethics of Sexual Difference, 20.
[59] Irigaray, An Ethics of Sexual Difference, 22.
[60] Irigaray, An Ethics of Sexual Difference, 22.

and confusion in Socrates, which I perceive as an intermediary state facilitating his transition from one conviction to another. This transitional phase enables him to relinquish his 'already established truths,'[61] by eliciting a pause in the conversation. In her book *Unlearning with Hannah Arendt*, Marie Luise Knott analyses Arendt's use of laughter. Knott emphasises that laughter can manifest in two distinct forms: One characterised by aggression and confined within conventional thinking and pre-existing knowledge, while the other creates a momentary pause that unravels the certainty of conclusions.[62] To illustrate how laughter is 'physically dependent on the ability to let go,' Knott draws upon Kant's description of laughter as the 'salubrious movement of the diaphragm.'[63] This physical and emotional release inherent in laughter momentarily disrupts the customary flow of conversation, interrupting established patterns of thought, meaning, and intelligence. By introducing a pause into conversation, laughter challenges the ordinary and paves the way for fresh perspectives and new ways of understanding to emerge.

Through Irigaray's interpretation, I perceive Diotima's laughter as a brief intermission, affording both her and Socrates the opportunity to gather their intellectual composure.[64] It nurtures an environment in which Socrates can comfortably scrutinise his convictions, facilitated by the ethical exchange she establishes. This setting places Socrates in an intermediate state, hovering between truth and falsity, as it induces him to pause and withhold judgement. This mirrors Knott's examination of Arendt's laughter, which she regards as a strategy of 'unlearning' which prompts an intellectual awakening.[65] In Irigaray's interpretation, Diotima's laughter serves as a response to Socrates' inability to grasp 'the existence or the in-stance of that which stands *between*.'[66] It cleverly rebuts what she perceives as Socrates' nonsensical assertions, leaving him unsettled and humbled in response. This reaction triggers a pause or interval, prompting Socrates to reconsider the

[61] Irigaray, *An Ethics of Sexual Difference*, 22. Again I must acknowledge how Socrates has now become the recipient of the pedagogical approach he has taken with others. However in Diotima's case, she has modified the approach to suit her relational focus with an educative function.

[62] Knott, Marie L (2013) *Unlearning with Hannah Arendt*, David Dollenmayer, trans, Other Press, 18, 21.

[63] Knott, Unlearning with Hannah Arendt, 21.

[64] Knott, Unlearning with Hannah Arendt, 19.

[65] Knott, Unlearning with Hannah Arendt, xi.

[66] Irigaray, An Ethics of Sexual Difference, 21.

statement that had triggered Diotima's laughter. Consequently, her laughter acts as a catalyst for Socrates 'to unlearn [his] dominant philosophical and cultural prejudices.'[67] In this instance, it enabled him to unlearn his previously held belief that love is a god.[68] Therefore, by establishing an ethical exchange between the parties, the laughter in Diotima's speech nurtures an intermediary space which disrupts established truths.

By cultivating this transitional state which dismantles certainty, laughter encourages a present connection with others and the world. This connection is achieved not through opposition, but rather through an open approach. As previously mentioned, Diotima's laughter, characterised by its teasing rather than humiliating tone, enables her to refrain from assuming a position of mastery over Socrates. Instead of cutting him off to advance her own conclusion, she grants him the space to continue his train of thought, facilitating an interaction that allows them to genuinely encounter each other. In her analysis of Arendt's use of laughter, Knott discusses how laughter can serve as a unifying force. She writes:

> When the partners in a debate concentrate only on their differences, identifying and insisting on them, they are emphasising what divides them, thereby letting the divide grow wider, gain significance, and become more palpable. By contrast, *laughter builds bridges* [...] difference and the experience of it are allowed to float free and feel secure in that hovering state [emphasis added].[69]

Here, Knott characterises laughter as having a unifying effect because it temporarily eases tensions and divisions. It momentarily suspends one's fixation on differences, creating a sense of connection and shared experience. This effect is reflected in Irigaray's reading as well. When Diotima laughs at Socrates, inducing a momentary pause and hesitation in their conversation, it prompts him to acknowledge the existence of perspectives which differ from his own. This pause prevents Socrates from continuing his metaphysical grappling within the confines of his own perspective and potentially missing the alternate viewpoint Diotima is trying to

[67] Boulous Walker, 'The Laughter of Hannah Arendt.'

[68] Plato, *The Symposium*, 80.

[69] Knott, Unlearning with Hannah Arendt, 14.

convey. Diotima's laughter serves as an interval during which Socrates can recognise the other and become more open to otherness.[70]

Importantly, her laughter is not merely a tool for encountering the other; it serves as a means to establish an *ethical* point of contact with the other. In addition to exposing Socrates to the uniqueness and difference of the other, Diotima encourages ethical engagement by nurturing an environment in which Socrates is receptive and open to being 'transformed by the encounter with the other.'[71] Her laughter effectively counters any attempt by Socrates to reduce or assimilate her perspective into his own. Instead, it opens Socrates to the possibility of reconsidering what he once deemed certain, driven by his receptiveness to the other as unknown. This encourages a more inclusive and diverse understanding of reality. This proves advantageous for Socrates as the educative effects of Diotima's laughter aids him in his engagement of philosophy in his quest for wisdom. Thus, in Irigaray's reading, Diotima's laughter is a tool for an ethical opening toward the other, thereby providing an ethical orientation for engaging in philosophy.

Building on this, laughter's significance extends beyond its role in establishing an ethical point of contact with the other. The momentary pause triggered by laughter illuminates a philosophical approach that encapsulates the essence of plurality. As Boulous Walker aptly observes, 'laughter provides the pause or interval necessary for us to move forward.'[72] In my view, 'moving forward' entails breaking free from the constraints of excessive seriousness deeply entrenched in dogmatism. Knott articulates this idea when she writes: 'Laughter makes available … confidence in the human power of resistance— against ideology and terror, against obscurantism, repression, dogmatism, and despotism.'[73] This underscores the subversive nature of laughter, as it challenges rigid structures by highlighting their absurdity or inconsistency. Consequently, it encourages individuals to engage in questioning and critical thinking rather than passively accepting dogmas. By highlighting Diotima's

[70] I recognise that reflective pauses can be achieved through various means, including simple pauses in conversation. However, Diotima's incorporation of laughter adds a joyful dimension, making the experience more welcoming and engaging for Socrates.

[71] Boulous Walker, *Slow Philosophy*, 92.

[72] Boulous Walker, 'The Laughter of Hannah Arendt.'

[73] Knott, Unlearning with Hannah Arendt, 10.

pedagogy of questioning and laughter, Irigaray's reading emphasises the importance of questioning our own 'established truths' so that we do not become intellectually stagnant and resistant to change. Diotima fosters intellectual curiosity and exploration by calling everything into question. This relentless questioning challenges established norms and hierarchies concerning the concept of love. This approach suggests that there is more to discover than a single, absolute truth. It not only deepens Socrates' understanding of the subject matter but also highlights plurality as a central aspect of philosophical inquiry. Diotima's laughter serves as a bridge to others and their diverse perspectives, embodying the essence of plurality by welcoming a variety of voices and viewpoints. It encourages Socrates to consider alternative perspectives, prompting critical examination of any biases or prejudices he may hold. This promotes a shift away from binary thinking towards embracing nuance and ambiguity. Embracing such complexity is what constitutes ethics,[74] as it teaches us to respect and engage with the beliefs and perspectives of others. Diotima's laughter advocates for a philosophical approach rooted in plurality by prioritising openness to alternative ideas and solutions rather than rigid adherence to a singular, fixed worldview.

Lastly, the value of open-mindedness and the willingness to challenge established truths is reflected in Irigaray's own open-ended reading of Diotima's speech. As previously mentioned, Irigaray refrains from critiquing the logical inconsistencies in Diotima's argument. Instead, she views these ambiguities and tensions as representations of the multiple voices which emerge from the text.[75] In line with Boulous Walker's perspective, 'there is, simply, no singular Diotima for Irigaray.'[76] Consequently, Irigaray concludes her reading in an open-ended manner, resisting the urge to align with one or the other of the opposing poles within Diotima's speech. She avoids attributing a singular conclusion to Diotima and, instead, invites readers to revisit Diotima's speech from the perspective of beauty rather than eros. Irigaray suggests that perhaps we have not adequately explored the category of beauty, leaving it relatively uncharted, and she prompts us to contemplate the untapped

[74] Boulous Walker, *Slow Philosophy*, 31.

[75] Irigaray, An Ethics of Sexual Difference, 27–29.

[76] Boulous Walker, *Slow Philosophy*, 89.

potential it might unveil. While a thorough exploration of this aspect might reveal the importance of rereading, as examined by Boulous Walker in her analysis in *Slow Philosophy*, within the context of the role of laughter, it serves as an illustrative example of what Irigaray identifies Diotima as doing: advocating for openness and the reconsideration of established truths. Just as Diotima's laughter does, Irigaray's ethical reading of Diotima's speech nurtures an approach to philosophy as a way of life with relational ethics at its core.

6. Conclusion

In Irigaray's reading of Diotima's speech, laughter offers an alternative way of thinking about what philosophy is and how to do it. Firstly, by interpreting Diotima's laughter as good-natured and aligning it with her pedagogical method of questioning, Irigaray illustrates how Diotima's laughter provides an escape route from the limitations of traditional hierarchical and oppositional approaches to philosophy. In doing so, Diotima establishes an ethical exchange where both participants are regarded as equals, fostering a more enriched and inclusive discourse, and cultivating a mutually beneficial learning process that denotes an openness to the other. Furthermore, Irigaray's reading showcases how laughter sketches a pathway of thought which disrupts established truths, remains open to otherness, and reflects the essence of plurality. By introducing a momentary pause in conversation, Diotima's laughter creates an intermediary state where new ways of understanding can emerge, free from the constraints of dogmatism. Coupled with Irigaray's own method of reading, which avoids a position of critique and acknowledges the nuances and open-endedness in Diotima's views, it becomes evident that laughter is a potent tool for instilling a philosophical approach with ethics at its core.

References

Arendt, Hannah (1978) The Life of the Mind, Secker & Warburg.

Boulous Walker, Michelle (2016) *Slow Philosophy: Reading against the Institution*, Bloomsbury Publishing.

Boulous Walker, Michelle (2021) 'The Laughter of Hannah Arendt,' *ABC Religion and Ethics*. https://www.abc.net.au/religion/the-laughter-of-hannah-arendt/13401584.

Chanter, Tina (2016) *Ethics of Eros: Irigaray's Rewriting of the Philosophers*, Routledge.

Irigaray, Luce (1993) *An Ethics of Sexual Difference*, Carolyn Burke and Gillian C Gill, trans. Cornell University Press.

Jones, Rachel (2011) *Irigaray*, Polity Press, ProQuest Ebook Central.

Knott, Marie Luise (2013) *Unlearning with Hannah Arendt*, David Dollenmayer, trans. Other Press.

Morreall, John (1982) 'A New Theory of Laughter,' *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* **42**: 243–54. doi:10.1007/bf00374037.

Plato (1967) *The Symposium*, Walter Hamilton, trans. Penguin Books.

Plato (1978) *The Collected Dialogues of Plato*, Edith Hamilton and Huntington Cairns, trans. Princeton University Press.

Reeve, C D C (2023) 'Plato on Friendship and Eros,' in Edward N Zalta and Uri Nodelman, eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2023/entries/plato-friendship/.

Optimism for the Value of Philosophy under Equilibrism

Johnny Kennedy[77]

Cambridge University

Abstract

In light of philosophical scepticism (scepticism about the possibility of philosophical knowledge), Beebee (2018) offers *equilibrism* as an alternative to knowledge as a conception of the aim of philosophy. This axiological thesis allows the philosophical sceptic to avoid metaphilosophical pessimism: the thesis that philosophy does not progress. However, in this paper, I scrutinise the value of philosophical work as it is conceived under *equilibrism*. I raise the 'Challenge from the Epistemic and Pragmatic Inadequacy of *Equilibrist* Philosophy' in order to emphasise the requirement for *equilibrism* to demonstrate the motivations for philosophical work as conceived under *equilibrism*. In response to this challenge, I locate two central features of *equilibrist* philosophical work (critique and formulating *equilibria*), and the epistemic and practical benefits they each confer, to defend an optimism about the value of philosophical work as conceived under *equilibrism*.

1. Introduction: Philosophical Scepticism and Equilibrism

[77] Johnny Kennedy is commencing his studies at the University of Cambridge, reading an MPhil in the History and Philosophy of Science and Medicine. He researches scientific realism, the philosophy of literature, and the interdisciplinary intersection of literature and science.

What is philosophy good for? For many naturalistic thinkers today, not much. We are familiar with the infamous opening lines of Stephen Hawking's The Grand Design:

> Philosophy is dead. [...] Scientists have become the bearers of the torch of discovery in our quest for knowledge.[78]

The Nobel laureate Francis Crick makes a similar gesture toward the displacement of philosophy's value by science:

> Essentially philosophers often ask good questions, but they have no techniques for getting the answers. Therefore you should not pay too much attention to their discussions. And we can ask what progress they have made. A lot of problems which were once regarded as philosophical, such as what is an atom, are now regarded as part of physics. Some people have argued that the main purpose of a philosopher is to deal with the unsolved problems, but the problems eventually get solved, and they get solved in a scientific way. If you ask how many cases in the past has a philosopher been successful at solving a problem, as far as we can say there are no such cases.[79]

The inadequacy of philosophical methods—what Crick calls 'techniques'—has been referred to by Beebee as support for the metaphilosophical position she calls 'philosophical scepticism'.[80] Beebee appeals to this inadequacy, along with the inadequacy of philosophical data and the systematic peer disagreement that is widespread throughout the discipline, in order to sceptically rebut (what she takes to be) the widely held assumption that philosophers can know any of the substantive philosophical claims that they make or presume about the world. If one believes—or has hitherto presumed—that philosophy is predominantly good for knowing about the world, then following Beebee to her sceptical conclusion may incite despair.

Specifically, we might make two distinct, despairing inferences:

[78] Hawking, Stephen (2010) *The Grand Design*, Bantam Books, 5, quoted in Pigliucci, Massimo (2022) 'Scientism and Liberal Naturalism', in M. De Caro & D. Macarthur, eds, *The Routledge Handbook of Liberal Naturalism*, Routledge, 374.

[79] Blackmore, Susan (2005) *Conversations on Consciousness*. Oxford University Press, quoted in Stoljar, Daniel (2017) *Philosophical progress: In Defence of a Reasonable Optimism*, Oxford University Press, xvii.

[80] Beebee, Helen (2018) 'Philosophical Scepticism and the Aims of Philosophy'. *Proceedings of the Aristotelian Society*, **118 (1)**, 1.

(1) if philosophy cannot produce knowledge, then philosophy cannot make progress;

AND

(2) if philosophy cannot produce knowledge, then philosophy has no value.[81]

Before we follow van Inwagen into pessimism about the 'futility of philosophy',[82] or go so far as to abandon philosophy altogether in favour of the more successful techniques of science, we ought to scrutinise the position that philosophical progress and value should be evaluated by philosophy's ability to produce knowledge. As we discover in this paper, our anxious reflection on philosophy's progress and value is directed by the shadow cast down from the incredible success of science in its ability to create consensus and use independent data to find patterns in nature.

As recognised by Dellsén et al., philosophers' optimism or pessimism about philosophical progress is often merely dictated by the standard one uses to evaluate philosophical success.[83] Appropriating standards of progress used in the philosophy of science, Dellsén et al. introduce three other possible candidates that could account for philosophical progress: a truthlikeness account, a problem-solving account, and their noetic account.[84]

Drawing on Lewis' remarks on a 'reasonable goal' for philosophy,[85] Beebee develops and defends her own alternative conception of philosophy's aim, what she calls 'equilibrism'. For Beebee, whilst no philosophical theory is ever going to achieve philosophical knowledge, we can rule out many philosophical theories and collate a collection of defensible (although inevitably underdetermined) theories. On this view, philosophy aims toward establishing an 'equilibrium', whereby all the indefensible philosophical theories have been discarded and we have finalised the

[81] These are distinct inferences that seem to be conflated by many commentators. We may vindicate the value of philosophy, and have a concept of philosophical success, without any concept of philosophical progress. This is the view presented in Shan, Yafeng (2022) 'Philosophy doesn't need a Concept of Progress', *Metaphilosophy*, **53(2-3)**.

[82] van Inwagen, Peter (1996) 'Review of Problems in Philosophy: The Limits of Inquiry by Colin McGinn', *The Philosophical Review*, **105(2)**, 253.

[83] Dellsén, Finnur, Lawler, Insa, & Norton, James. (2021). 'Thinking about Progress: From Science to Philosophy'. *Noûs*.

[84] Dellsén et al., 14-18.

[85] Lewis, David (1983) *Philosophical Papers, Volume I.* Oxford University Press, x, quoted in Beebee, 15-16.

collection of philosophical views that can withstand philosophical examination. For example, philosophy ought not to hope to produce knowledge about free will, but it can create alternative, competing theories about free will, and, subsequently, distinguish the weak theories from our best. In light of Beebee's view that philosophy cannot produce knowledge, equilibrism—the point at which our best competing theories for a given phenomena are refined—is taken to be an achievable aim that can also be used as a standard for measuring philosophical progress.[86] In this way, Beebee's equilibrism offers a salve to soothe the pangs of the first despairing inference I listed above.

In this paper, I will take it for granted that equilibrism provides an adequate and suitable standard upon which to measure philosophical progress. Instead, I will focus on scrutinising the adequacy of equilibrism as an explanation of the value of philosophy, and whether the equilibrist can diffuse the second despairing inference listed above. Does the equilibrist conception of philosophy help us make sense of the philosophy's value; a value that the distinguished Hawking and Crick cannot see?

In §2, I will formulate a challenge as to the value of philosophy as conceived under equilibrism: Challenge from the Epistemic and Pragmatic Inadequacy of Equilibrist Philosophy. This challenge contends that 'equilibrist philosophy' (philosophy as conceived under the aim equilibrism describes) can only tell us about different theories that we create, but it cannot help us know or understand anything about the world, nor does it have any practical value. In §3-4, I respond to this challenge with two justifications for the epistemic and practical value of equilibrist philosophical work. Equilibrist philosophy operates in two ways.[87] Firstly, equilibrist philosophy places theories under scrutiny. I raise The Argument from the Inevitability of Philosophical Views (§3.1) in order to demonstrate how this critical role relieves us of philosophical blunders that we would otherwise inevitably make. Secondly, philosophy constructs alternative and varied theories about a given phenomena. Following Catherine Elgin's conception of understanding, I claim that by scrutinising and producing the best possible competing theories for a given phenomena, philosophy allows us to conceive how the world might reasonabley be

[86] Beebee, 15.
[87] These ways are related, as elaborated in §3.1.

taken to be. Moreover, I offer that although this understanding might not have an obvious pragmatic value, it confers the practical benefit of better decision-making on issues that relate to philosophical subject matters.

2. Why Would the Equilibrist Keep Doing Philosophy?

2.1 Philosophy under Equilibrism

Before we articulate the challenge regarding the value of equilibrist philosophy, it is important to clarify the equilibrist picture of philosophical work. On a strict interpretation of philosophical scepticism, philosophical methodology will never be able to determine the justification for one particular philosophical theory for a given subject.[88] Two central motivations for this stance are: (1) the inadequacy of the data, and (2) the different weight philosophers assign to the theoretical virtues. On this view, philosophical methodology faces a serious underdetermination problem, and whilst philosophical examination may be able to refute indefensible philosophical views, there will always be multiple views that could reasonably be defended. Equilibrism contends that the aim of philosophy is to 'find out what equilibria there are that can withstand examination',[89] where the equilibria are those theories that can reasonably be defended.

Beebee recognises that under an equilibrist conception of philosophy, philosophers may be able to better recognise that some arguments are unproductively intractable because they are working with different data, or weigh the theoretical virtues differently. For example, one significant portion of our philosophical data is our intuitions, which may not have the feature of independence in the way that empirical data does for scientific theories. As recognised by van Fraassen, empirical success in the sciences, and the consequent formation of consensus for successful scientific theories, is produced by science's methodological commitment to independent data.[90]

[88] A fallible philosophical scepticism might be preferable. It isn't clear why Beebee would have to insist that it is impossible that, on a rare occasion, our intuitions and assessment of the weight of the virtues could align for a particular topic. For example, Goedel's theorems of incompleteness are two theories in logic that have widespread consensus and seem to have been accepted as philosophical knowledge for half a century. This discussion, however, falls outside the scope of this paper.

[89] Lewis, x, quoted in Beebee, 15-16.

[90] van Fraassen, Bas C. (2002) *The Empirical Stance*, Yale University Press, 159.

In science, independent data is that which can be observed by anybody (at least anybody with the requisite scientific training and background to comprehend the observation). For example, Perrin was able to conduct fruitful experiments into Brownian motion (the movement of tiny granules suspended in water), and sway the scientific community into consensus concerning the existence of atoms, due to his measurement of data specified by a demarcated parameter: the mean kinetic energy of the granules.[91] By measuring the behaviour of these granules on this particular parameter, Perrin was able to confirm Einstein's predictions of the mean displacement of the granules and their rotation energy,[92] and, consequently, cite thirteen different ways of precisely calculating the same quantity for Avogadro's number (N); the number of particles in a unit known as a mole (mol).[93] The measurements relating to the granules' mean kinetic energy are independent to the extent that they could have been observed by any of Perrin's trained colleagues had they wanted to follow Perrin around the laboratory, checking Perrin's microscopes and taking their own photographs. In this sense, these measurements are independent data.

For Beebee, in philosophy we generally find a different story associated with the appeal to data. A philosopher may invite their colleague to recognise an intuition by means of a thought experiment. However, as we are well aware, oftentimes their trained colleague looks through the metaphorical "lens" of the thought experiment, only to make a completely contradictory observation. Of course, if philosophy is aimed toward knowledge, then the philosophers can only ensure the reliability of the data by disputing what each other "observe" about their own intuitions, and concoct tactics to compel each other to see the intuition which they simply do not have. Equilibrism provides philosophers with a conception of their activity that relieves them of the necessity for further table-thumping and foot-stomping disputes where

[91] Psillos, Stathis (2011) 'Moving Molecules above the Scientific Horizon: On Perrin's Case for Realism', *Journal for General Philosophy of Science*, **42(2)**, 353; Chalmers, Alan (2020) 'Drawing Philosophical Lessons from Perrin's Experiments on Brownian Motion: A Response to van Fraassen'. *The British Journal for the Philosophy of Science*, **62**, 722.

[92] Einstein, Albert (1905) 'On the Motion of Small Particles Suspended in Liquids at Rest Required by the Molecular-Kinetic Theory of Heat', in A. D. Fürth, ed, *Investigations on the Theory of the Brownian Movement*, Dover Publications, 17.

[93] Psillos, Stathis (1999) *Scientific Realism: How Science Tracks Truth*. Routledge, 19.

it is no longer fruitful for the equilibrist aim. However, whilst this is a good motivation for preferring an equilibrist conception of the aim of philosophy over philosophy conceived under the aim of knowledge, I will now raise the problem of understanding why an equilibrist philosopher would bother undertaking philosophical work at all.

2.2 Challenge from the Epistemic and Pragmatic Inadequacy of Equilibrist Philosophy

One feature of philosophy, as it is conceived under equilibrism, is that philosophers ought not to believe the philosophical views that they accept and commit to. Beebee, a metaphilosophical sceptic, appropriates van Fraassen's anti-realist account of scientific acceptance as an alternative to belief. For Beebee, the philosopher can accept a certain philosophical theory about free will without actually believing in the theory. Here, acceptance of a theory merely amounts to a pragmatic commitment to:

1) Confront any future phenomena by means of the conceptual resources of this theory,
2) Be willing to answer questions ex cathedra, and
3) Assume the role of the explainer.

By accepting a theory 'one commits to speak and write and act as though the theory is true',[94] without believing it is so. Beebee contends that:

> This attitude can be applied to the working philosopher no less than to the working scientists.[95]

Beebee, however, overlooks one significant feature of van Fraassen's account of the acceptance of scientific theories.

On van Fraassen's account, our acceptance of scientific theories is pragmatically connected to practical concerns. For van Fraassen, even though scientific theories are not taken to be completely true, he does acknowledge that science uncovers the 'actual regularities' that are to be found in nature.[96] Practically equipped with

[94] Beebee, 21.

[95] Beebee.

[96] van Fraassen, Bas C. (1980) *The Scientific Image*. Oxford University Press, 40.

knowledge of these regularities, we can predict and thereby manipulate phenomena in order to develop technologies, take photos on Mars, and immunise ourselves against certain viruses. Therefore, even on van Fraassen's anti-realist interpretation of scientific activity, our scientific acceptance can be justified with reference to practical concerns.

On the other hand, what motivation is there for an equilibrist philosopher to accept their philosophical theories in the way that the scientific anti-realist accepts scientific theories? If, on the equilibrist account, philosophers do not believe any philosophical theory, then what is the point of philosophical work? As discussed above, van Fraassen gives reasons as to why a scientist is rightly still willing to commit herself to a theory. It is not immediately obvious what motivates a philosopher to commit themself to a theory. Does equilibrist philosophical activity have either any epistemic value (for understanding the world) or any practical value (for helping us navigate our way through life)? This is the Challenge from the Epistemic and Pragmatic Inadequacy of Equilibrist Philosophy. If Beebee's thesis does not capture the value of philosophy, then this may be grounds for resisting equilibrism as an adequate conceptualisation of philosophy's aim.

Equilibrist philosophy needs a justificatory story if we are going to conceive of philosophy as a worthwhile exercise. Whilst equilibrism might equip philosophers with a response to pessimism about philosophical progress, they aren't equipped to respond to those naturalistic thinkers, in the vein of Hawking and Crick, who might continue to assert that philosophy is dead and has nothing to really offer us in the wake of the success of science. In §3-4, I will give two motivations for optimism about the value of equilibrist philosophical work.

3. Equilibrist Philosophy and the Value of Critique

Recalling the account of equilibrism offered in §1.1, we can identify two central features of equilibrist philosophy:

1) the process of examining and critiquing philosophical views; and

2) the process of developing and improving philosophical views as possible candidates for equilibria.

In the following two sections, I will expand on these two features, and reveal how they can direct us toward understanding how philosophy is valuable and worthwhile, even if it does not produce knowledge. Hopefully, by the end of Sections III and IV, the equilibrist will be equipped to address the Challenge from the Epistemic and Pragmatic Inadequacy of Equilibrist Philosophy.

3.1 The Argument from the Inevitability of Philosophical Views

Philosophical views are not as dispensable as the comments from Hawking and Crick imply. Philosophical commentators on these passages are often quick to recognise that both Crick and Hawking are themselves expounding philosophical views in their critique of philosophy.[97] Pigliucci is right to suggest that Hawking's dismissal of philosophy's adequacy can be seen as self-defeating when Hawking's entire The Grand Design (2010) is itself 'best characterised as a popular treatise on the philosophy of cosmology'.[98]

Although we might not notice it, we take on philosophical views all the time. Moreover, our beliefs and actions often imply certain philosophical assumptions which may not be consciously held. Alternatively, we may hold beliefs that on their face don't seem philosophical which have controversial philosophical corollaries. As recognised in the philosophy of science, if we believe that the entities described by our best scientific theories are true, which very many people do (whether or not they have philosophically deliberated about it), then we are implicated in the belief that the theoretical criteria for theory preference—the theoretical virtues—are in some sense 'truth conducive'. The view that a simpler theory is a better candidate for truth than a more complex theory (ceteris paribus) seems difficult to justify without some underlying metaphysical presupposition (or faith) that the world is, in some sense, simple.[99] In this way, what may seem like ordinary views to many people are either themselves philosophical, or they entail ones that are philosophical.

[97] See Pigliucci 2019, 374 and Stoljar 2017, 3.
[98] Pigliucci 2019, 374.
[99] See Bueno 2015, 674.

Not only do we find ourselves assuming, implying, or taking on philosophical views in our ordinary lives (let alone in crucial moments of our lives), but we can also see how the human intellect has a positive tendency to interpret the world by means of philosophical theories and explanations. As noted by van Fraassen, the history of philosophy reveals the human impulse to explain empricial phenomena by reference to some further, theoretical entity. It seems human beings cannot merely accept what they experience at face value without giving some account of it. For van Fraassen, we specifically go astray when we insist on an inference method he refers to as 'explanations that proceed by postulation'.[100] This is the inference method of explaining a phenomenon by appealing to the reality of certain entities or aspects of the world not already evident in experience. On van Fraassen's account of empiricism, the empiricist philosopher stands in negative opposition to this 'theoretical tendency':

> That is why we are ready to call Aristotle more of an empiricist than Plato and speak of an empiricist turn at that point. Aristotle called Plato's followers back from high theory to empirical inquiry. That is also why we think of the late fourteenth-century nominalists as the parents of British empiricism: they staged a rebellion against an Aristotelian tradition that had wandered far away from Aristotle's empirical focus … Similarly, the [logical] positivists and later empiricists staged yet another new beginning for empiricism, in their critical opposition to the metaphysics of their day.[101]

In The Gay Science, Nietzsche provides a comparable commentary on the history of philosophy as a flight from one's actual life (and experience of the world) toward another theoretically constructed world. On Nietzsche's assessment, Plato is unable to accept the world in which we live, the 'rerum concordia discors' (discordant concord of things).[102] Following Nietzsche's provocative commentary, Plato constructs the theoretical world of the forms, a theoretical world that explains the world of appearances and purports to be more fundamental to it, as a way of escaping from and 'denying' the reality of the world of appearances. Across the

[100] van Fraassen 2002, 37.

[101] van Fraassen 2002, 36.

[102] Nietzsche, Friedrich (1887) *The Gay Science*, W. Kaufmann, trans, Vintage, 30.

history of human thought (as thought of by Nietzsche) we fall prey to a metaphysical impulse, whether it is the divine world contemplated by Christians or the objective world of laws described by science. This is the impulse—a 'metaphysical need'[103]—to construct a theoretical explanation of what we experience.

The indispensability of philosophy, and the human tendency to fall prey to theoretical thinking, calls for philosophical examination of those philosophical views that we find ourselves accepting or implicating ourselves in. Philosophical examination involves revealing the philosophical assumptions or implications of certain scientific, political, religious (or other ordinary) beliefs that we might hold or entertain. It also involves evaluating those philosophical views and determining whether they are outright untenable, or whether they are defensible. This is completely commensurate with the equilibrist vision of philosophy offered by Beebee.

In this fashion, equilibrist philosophy is valuable as a way of relieving us of indefensible philosophical views we might unwittingly accept or imply. Of course, philosophy already performs this function. We know this with reference to our own life as philosophers. We have disembarrassed ourselves of disastrous philosophical blunders due to our study of philosophy and uncovered dubious philosophical assumptions that underpinned the way that we were thinking. Hopefully, we have also helped relieve others of the burden of an unexamined philosophical assumption, even if only in our personal (as opposed to academic) lives. The Argument from the Inevitability of Philosophical Views shows that philosophy is valuable, even essential, for examining the views that we inevitably take or implicate ourselves in. Even if philosophy dispenses with knowledge as its aim, it has a critical value: one that dismissive remarks like Hawking's and Crick's overlook.

3.2 Philosophy as the Discipline of Critique

[103] Nietzsche, 131.

The defence of equilibrist philosophy has revealed a distinctive feature of philosophy as a discipline. This feature is its commitment to critiquing and examining philosophical views. Of course, one might respond by pointing out that philosophy is not the only discipline that is willing to take on a critical attitude. Literary critics scrupulously dissect each others' assessments in order to expose a misreading or an oversight. Experimental scientists will expend vast quantities of resources to meticulously construct elaborate or intricate experimental apparatus, just so they can assess whether the theories of their theoretical colleagues stand up to the test.[104] However, following Priest, philosophy is distinctive for its 'unbridled' willingness to examine and critique any view whatsoever:

> Anything is a fit topic for critical scrutiny and potential rejection [...] even the efficacy of critical reasoning itself.[105]

The philosophical interlocutors of Beebee's paper (Argle, Bargle, Cargle, Dargle, Fargle) are absorbed in seeking ways to scrutinise each others' views about various different topics. In these discussions, there may be no shared theoretical bedrock that the interlocutors share: nothing is taken for granted. What's more, as emphasised by Beebee, they might not even agree on what makes a theory good, weighing the value of different theoretical virtues differently to each other.[106] Philosophy's insistence on examining every element of each other's views is one distinctive feature of philosophical critique. Priest recognises that although scientists are encouraged to scrutinise and test novel theories, ideas, and results, 'no one is encouraged to question well entrenched and established parts of the scientific corpus'.[107]

Following Kuhn, science's success and efficiency is facilities by an element of dogmatism. For Kuhn, the dogmatic element of science is typified by the scientific textbook.[108] Science is not taught critically. However, due to this pedagogical dogmatism, science is able to produce specialists quickly and efficiently because they have been trained to work within an underlying general theory, or, in Kuhnian terms, a 'paradigm'. Scientists are not trained to question the general theory, but to

[104] Priest, Graham (2006) 'What is philosophy? Philosophy', **81(2)**, 201.
[105] Priest, 201.
[106] Beebee, 8.
[107] Priest, 201-02.
[108] Kuhn, Thomas (1962) *The Structure of Scientific Revolutions*, The University of Chicago Press, 136-40.

try to solve its specific problems. In Shapin's sociological account of the intricate networks of trust upon which science operates, he emphasises that to take a sceptical attitude to the presumptions that underpin normal (specialist) experimental science would be enormously costly and time-consuming. As Shapin comically recognises, one would have to set up "counter-laboratories" to negatively match each laboratory we have today.[109]

However, one must acknowledge that science will, from time to time, challenge the well-entrenched foundations within its paradigm. However, these moments of conceptual revolution are exactly where the boundaries between philosophy and science begin to blur. Priest recognises that when a scientist engages in critiques 'that go beyond the bounds of what is normally permitted, they are engaging in philosophy'.[110] This view is commensurate with Kuhn's characterisation of scientific revolutions, where research transitions from 'normal' science to 'extraordinary' research, and scientists must have 'recourse to philosophy and [...] debate over fundamentals'.[111]

Philosophy has a distinctive role in our epistemic projects. It is the discipline where any view is liable to receive examination. Whether in religion, politics, or features of our social interactions, philosophy has been distinctively invaluable for scrutinising views that we take for granted, or wouldn't think to countenance rejecting.

4. The Value of Equilibria

The positive counterpart—to the negative value of philosophical critique outlined in §3—is the value of collating candidates for defensible equilibria. An important part of philosophical critique is the creative construction of alternative or improved versions of a view. In this final Section IV

, I will introduce the importance of the constructive aspect of equilibrist philosophy by demonstrating its relationship with the critical aspect of philosophy (§4.1). Appropriating discussions of understanding from Elgin's recent True Enough,[112] I

[109] Shapin, Steven (1994) *A Social History of Truth: Civility and Science in Seventeenth-Century England*, University of Chicago Press, 19.

[110] Priest, 202.

[111] Kuhn, 91.

[112] Elgin, Catherine (2017). *True Enough*, MIT Press.

will then argue that equilibrist philosophy offers us the value of non-factive "modal understanding" of its subject matter (§4.2).

4.1 Critique and Construction

It is important to recognise that the two features of equilibrist philosophy I have located— its critical and its constructive—are not totally distinct from each other. Critique is most forceful when it is coupled with the proposal of a competing theory.

Every view has its problems. This is true for philosophical views as well as views generally, including scientific views. As recognised by Larry Laudan:

> Almost every [scientific] theory in history has had some anomalies or refuting instances; indeed, no one has ever been able to point to a single major theory which did not exhibit some anomalies.[113]

Even our best contemporary science is known to have serious problems which it must overcome.[114]

Let's consider, for instance, the rotation problem for spiral galaxies as an example of an inconsistency in our best science. There is a known contradiction between the predictions of how spiral galaxies rotate under Newtonian gravitational theory ('NTG') and what we actually observe of spiral galaxies.[115] Assuming that galaxies have a greater concentration of mass as you move toward the centre (which is indicated by astronomical observations), then the centre will spin faster than the spiral arms, and there will be a decline in the radial speed as you move away from the centre. This means that we ought to observe that the inside of the galaxy is spinning a lot faster than the outside. In 1959, it was discovered that the Triangulum Galaxy, M33, did not exhibit the decline in radial speed predicted by NTG. M33's rotation curve was found to be flat: the outer part of the galaxy was spinning at much the same radial speed as the centre. However, this observation did not amount to a refutation of NTG. Rather than give up the theory, scientists are instead

[113] Laudan, Larry (1977) *Progress and its Problems: Towards a Theory of Scientific Growth*, Routledge, 27.

[114] Consider the winners of the 2022 Nobel Prize for Physics—Alain Aspect, John Clauser and Anton Zeilinger—whose experimental work has highlighted the inconsistencies between our best theory of space and time (relativity theory) and our best theory of particulate matter. See, for an overview, Aspect, Alain (2015) 'Closing the Door on Einstein and Bohr's Quantum Debate'. *Physics*, **8**, 123.

[115] See Colyvan, Mark (2008) The Ontological Commitments of Inconsistent Theories. Philosophical Studies, **141**, 166.

experimentally looking for evidence for dark matter, a theoretical entity that would account for the contradictions between the observations and NTG.

As recognised by Laudan, issues with a theory are only taken to be devastating for a theory when they are made in conjunction with a positive, and competing alternative. For example, the perihelion precession of Mercury was known to be inconsistent with Newtonian mechanics well before Einstein's relativity theory. However, when Einstein developed relativity theory, he appealed to the perihelion precession as an observation that classical mechanics could not account for, one that is consistent with his competing candidate (general relativity). This is one of the central pieces of evidence that were decisive for the rejection of classical mechanics in favour of relativity theory.

It is for this reason that Laudan generalises:

> Unsolved problems … count as genuine problems only when they are no longer unsolved. Until solved by some theory in a domain they are generally only "potential" problems rather than actual ones.[116]

Critique finds its force with the construction of theoretical alternatives, one that can better account for the problems of the accepted theory. This is a feature of how philosophers can provide compelling critiques. Bargle claims that Argle's theory of holes is inconsistent with common sense, and this critique really gets Argle's attention because Bargle has formulated a competing, alternative theory of holes that is commensurate with common sense. Similarly, in the philosophy of science, contemporary resistance to scientific realism is not simply due to the issues with Inference to the Best Explanation and the 'No Miracles Argument' raised by van Fraassen,[117] but also his development of a strong alternative that can account for those limits: constructive empiricism.

4.2 Equilibria and Understanding

> Possibility is higher than actuality.

[116] Laudan, 18.

[117] van Fraassen (1980), 19-22.

— Heidegger, Being and Time.[118]

Equilibria are not only important for their role in giving force to philosophical critique, but I will also offer an argument in this final subsection that they give us a positive epistemic understanding (as opposed to knowledge) of their subject matter. This argument relies upon a non-factive view of understanding. That is, the view that understanding is a kind of cognitive achievement that cannot be merely reduced to knowledge. Of course, if the understanding were reducible to knowledge, the equilibrist would have to concede that philosophy cannot give us understanding of its subject matter.

Supposing the adequacy of the characterisation of knowledge as justified, true belief (with the exceptions being those of the kind raised by Gettier),[119] philosophers have defended the reductionist view that understanding can only be thought of as a justified, true belief about a certain subject matter. Following Aristotle, we might say we understand a given subject matter merely when we have knowledge of its causes. In this way, understanding a phenomenon is a certain, specific kind of knowledge.

However, these accounts overlook how we also use the concept of understanding to refer to a 'non-factive' cognitive achievement. Following Catherine Elgin, we need not believe true propositions in order to have an understanding of a given subject matter. For Elgin, this is typified by how we conceive of scientific understanding, which regularly uses 'idealisations' as vehicles for grasping a subject matter that would be otherwise difficult to 'grasp' by merely referring to true propositions. For Elgin:

Models and idealizations are … more than heuristics. They are ineliminable and epistemically valuable components of the understanding science supplies.[120]

For example, light can be modelled as a wave or as a particle. Both models exemplify certain features of how light behaves, and are thereby helpful vehicles for grasping those features. In this way, they are helpful but not exactly true ways ('idealisations')

[118] Quoted in Sheehan, Thomas (1993) 'Reading a Life: Heidegger and Hard Times', in C. Guignon, ed, *The Cambridge Companion to Heidegger*, Cambridge University Press, 93.
[119] Gettier, Edmund (1963) 'Is Justified True Belief Knowledge?' in D. Pritchard & R. Neta, eds, *Arguing about Knowledge*, Routledge.
[120] Elgin, 1.

of conceiving of the phenomena of light. Understanding, in this sense, refers to more than merely knowing certain facts about light, it is a capacity for being able to locate and grasp various connections between the body of information (models of light) and the actual subject matter (light). Another compelling and comparable argument is developed by Ivanova, who recognises that

If truth is a necessary condition for understanding, it would follow that past scientists lacked understanding of phenomena for which they had advanced empirically successful (but from our perspective false) theories.[121]

By showing that we have a concept of the understanding that is not dependent on truth, and thereby not reducible to knowledge (where knowledge is defined as justified, true belief), understanding becomes a cognitive achievement which the equilibrist philosophy might hope to achieve.

Elgin refers to the concept of a 'tether' in order to refer to those connections between a body of information and the subject matter. As Elgin recognises, some bodies of information may not have any tethers to their subject matter whatsoever. For example,

> 'Even if astrology offers a comprehensive, internally coherent account of the cosmos, it yields no understanding because it lacks a suitable tether.'[122]

However, Elgin clarifies that we can understand astrology as a body of information to be understood. When we say 'Paul understands mythology', we are not referring to Paul's ability to see and use connections between mythology and actual historical events. Rather, we refer to Paul's account of mythology and his ability to locate its connections as it is tethered to mythology as a body of knowledge itself. In the same sense, we can have a better or worse understanding of astrology, and, in turn, a better or worse understanding of philosophy.

It is important to recognise that we want more than merely this kind of understanding out of philosophy. We don't engage with philosophy merely out of a historical interest of understanding what the body information, but with a hope or

[121] Ivanova, Milena (2020) 'Beauty, Truth, and Understanding', in Ivanova and French, eds, *The Aesthetics of Science*, 98, citing De Regt, Henk (2015) 'Scientific Understanding: Truth or Dare?', *Synthese*, **192**: 3781–3797.
[122] Elgin, 45.

confidence that it will give us some benefit or insight in our project of understanding and engaging with the world. In response to this, we ought to recognise that philosophy, as a body of knowledge, does not completely lack tethers to the world. These tethers are just of a different kind than we are used to in areas of knowledge (like the natural sciences). On the equilibrist account of philosophy, we cannot extrapolate any knowledge of the world from this body of information. We can, however, have an understanding of what philosophical views are indefensible, and, therefore, could not be rationally considered to have any tethers to the world. These are the views that fall by the wayside in the equilibrist project for philosophy.

On the other hand, we can also have an understanding, and locate, how our equilibria would be tethered to the world if they were true. In this sense, equilibrist philosophy can offer what we might call a 'modal understanding' of its subject matter: that is, an understanding of different ways of rationally conceiving how the phenomena might be. On this view, we ought to conceive of equilibria as tied to the world by modal tethers.

To illustrate with an example, let's suppose that we have reached philosophical 'equilibrium' with regard to the problem of free will. In this instance, we have rejected all those indefensible philosophical views on free will, and we have arrived at three different defensible equilibria. Those equilibrist philosophers have worked scrupulously in assessing, improving, and refuting all different kinds of views and arguments relating to free will. They know the surviving equilibria front-to-back, and they have all done their best at attempting to knock down at least one of those remaining equilibria. I contend that those equilibrist philosophers will have a valuable modal understanding of how the phenomena, 'free will', might reasonable be taken to be. This understanding is also one of practical importance. I'd claim that government and legal institutions ought to defer to these equilibrist philosophers' assessments in creating policy, legislation, or giving judicial decisions in matters relating to free will (or accountability, and responsibility). After all, those institutions may be basing their policy, law, or judgement on a view of free will that those equilibrist philosophers have shown to be disastrously indefensible. As insisted by Blackburn, if we don't have the appetite to engage with the problems of philosophy, we may choose to shrug them off.[123] However, he warns that 'difficulties have a way of biting back … while we don't know our way about, our practices will risk being muddled and unjust'.[124] Ofcourse, the equilibrist philosopher does not claim to know her way about, but she can tell us which paths are blunders, and which are defensible. By deferring to philosophers, institutions may be able to take into account those equilibria for free will, and make decisions that are sensitive to them. Even where these equilibria are diametrically opposed, the equilibrist philosopher can assist in developing a more balanced institutional or policy approach to a given issue where reasonable minds may simply disagree. In this sense, the 'modal understanding' offered by equilibrist philosophy offers important epistemic and practical value that vindicates the philosophical work done under an equilibrist conception of philosophy.

5. Conclusion

In §2.2, I noted that to make sense of philosophical 'acceptance' as developed by Beebee (in light of van Fraassen's concept of scientific acceptance), the equilibrist needs to show that philosophers will have a motivation to commit themselves to philosophical work as conceived under equilibrism. In this essay, I have endeavoured to answer the question: if philosophy must dispense with knowledge as its aim, then what motivates the value of philosophical work? In so doing, I have, in part, uncovered two significant aspects of philosophy's value. In §3, I discussed philosophy's value as critiquing those philosophical views we seem to inevitably find ourselves taking on or implicating ourselves in. In §4, I offered 'modal understanding' as a way of conceiving of the epistemic and practical value of equilibria, which need not be true to give us insight into the world, specifically, how the world might reasonably be taken to be.

[123] Blackburn, Simon (2006) *Truth: A Guide for the Perplexed*. Penguin. 112.
[124] Blackburn.

References

Aspect, Alain (2015) Closing the Door on Einstein and Bohr's Quantum Debate, *Physics*, **8**: 123. https://doi.org/10.1103/physics.8.123.

Beebee, Helen (2018) 'Philosophical Scepticism and the Aims of Philosophy', *Proceedings of the Aristotelian Society*, **118**: 1-24. https://doi.org/10.1093/arisoc/aox017.

Blackburn, Simon (2006) *Truth: A Guide for the Perplexed*. Penguin.

Blackmore, Susan (2005) *Conversations on Consciousness*. Oxford University Press.

Bueno, Otávio, & Shalkowski, Scott (2015) 'Modalism and Theoretical Virtues: Toward an Epistemology of Modality', *Philosophical Studies*, **172(3)**: 671-689. https://doi.org/10.1007/s11098-014-0327-7.

Chalmers, Alan (2020) 'Drawing Philosophical Lessons from Perrin's Experiments on Brownian Motion: A Response to van Fraassen', *The British Journal for the Philosophy of Science*, **62**: 711-732. https://doi.org/10.1093/bjps/axq039.

Colyvan, Mark (2008) 'The Ontological Commitments of Inconsistent Theories', *Philosophical Studies*, **141**: 115-123. https://doi.org/10.1007/s11098-008-9266-5.

Dellsén, Finnur, Lawler, Insa, & Norton, James (2021) 'Thinking about Progress: From Science to Philosophy', *Noûs*, 1-27. https://doi.org/10.1111/nous.12383.

De Regt, Henk (2015) 'Scientific Understanding: Truth or Dare?', *Synthese*, **192**: 3781–3797. https://doi.org/10.1007/s11229-014-0538-7

Elgin, Catherine (2017) *True Enough*. MIT Press.

Einstein, Albert (1905) 'On the Motion of Small Particles Suspended in Liquids at Rest Required by the Molecular-Kinetic Theory of Heat', in A. D. Fürth, ed, *Investigations on the Theory of the Brownian Movement:* 1-18. Dover Publications.

Gettier, Edmund (1963) 'Is Justified True Belief Knowledge?', in D. Pritchard & R. Neta, eds, *Arguing about Knowledge*: 14-15. Routledge.

Hannon M., & Nguyen J. (2021) Understanding Philosophy. Manuscript. https://philpapers.org/rec/HANUP.

Hawking, Stephen (2010). *The Grand Design*. Bantam Books.

Heidegger, Martin (1953). *Being and Time.* J. Stambaugh, trans, State University of New York.

Ivanova, Milena (2020) 'Beauty, Truth, and Understanding', in Ivanova and French, eds, *The Aesthetics of Science*: 86-103. Routledge.

Kuhn, Thomas (1962) *The Structure of Scientific Revolutions*. The University of Chicago Press.

Laudan, Larry (1977) *Progress and its Problems: Towards a Theory of Scientific Growth*. Routledge.

Lewis, David (1983) *Philosophical Papers, Volume I.* Oxford University Press.

Nietzsche, Friedrich (1887) *The Gay Science*. W. Kaufmann, trans, Vintage.

Pigliucci, Massimo (2022) 'Scientism and Liberal Naturalism', in M. De Caro & D. Macarthur, eds, *The Routledge Handbook of Liberal Naturalism:* 371-382. Routledge.

Priest, Graham (2006) 'What is philosophy?', *Philosophy*, ***81*****(2)**: 189-207. https://doi.org/10.1017/s0031819106316026.

Psillos, Stathis (1999) *Scientific Realism: How Science Tracks Truth*. Routledge.

Psillos, Stathis (2011) 'Moving Molecules above the Scientific Horizon: On Perrin's Case for Realism', *Journal for General Philosophy of Science*, **42(2)**: 339-363. https://doi.org/10.1007/s10838-011-9165-x.

Shan, Yafeng (2022) 'Philosophy doesn't need a Concept of Progress', *Metaphilosophy*, **53(2-3)**: 176-184. https://doi.org/10.1111/meta.12526.

Shapin, Steven (1994) *A Social History of Truth: Civility and Science in Seventeenth-Century England*. University of Chicago Press.

Sheehan, Thomas (1993) 'Reading a Life: Heidegger and Hard Times', in C. Guignon, ed, *The Cambridge Companion to Heidegger*: 70-96. Cambridge University Press.

Stoljar, Daniel (2017) *Philosophical Progress: In Defence of a Reasonable Optimism*. Oxford University Press.

van Fraassen, Bas C. (1980) *The Scientific Image*. Oxford University Press.

van Fraassen, Bas C. (2002) *The Empirical Stance*. Yale University Press.

van Inwagen, Peter (1996) 'Review of Problems in Philosophy: The Limits of Inquiry by Colin McGinn', *The Philosophical Review*, **105(2)**, 253–256. https://doi.org/10.2307/2185726.

"I'm the same – but I'm not": Transracial Adoptees, Hermeneutic Injustice, and Coalitional Politics

Beau Kent[125]

The University of Melbourne

Abstract

This paper aims to achieve two goals: first, to argue that transracial adoptees lack the critical resources to adequately articulate their experiences, which constitutes a hermeneutical injustice. Second, to point towards potential strategies or ways of thinking that could assist adoptees in navigating their experiences which are yet to be widely recognised, both individually and as a community. I will argue that there is a relationality to the adoptee identity which means that there are few conceptual resources that adoptees can draw on that capture their experience at the intersection of white enculturation and a body of colour; this constitutes a hermeneutical injustice. I then provide a potential method for concept generation using Mariana Ortega's notion of 'hometactics' to argue that one way forward may be to engage in a practical 'making-do' rather than try to create more theoretically rigorous and abstract concepts. Finally, I point towards the possibility of coalitional politics through the notion of complex communication in order to create strong political intra and inter-group alliances.

[125] Beau Kent (he/they) is a recent graduate (2023) from the honours philosophy program at the University of Melbourne. He completed a thesis on the phenomenology of transracial, transnational adoptees and the critical phenomenology of the Latina feminist tradition. His philosophical work centres predominantly around critical phenomenology, adoption studies, and deconstruction, but they also have an interest in analytic philosophy of language and social epistemology. Beau currently works as a research assistant at the Alfred Deakin Institute at Deakin University.

1. Introduction

The goal of this paper is to act as a prolegomenon to future work in the philosophy of transracial adoption. Therefore, this analysis will be predominantly exegetical. I wish to achieve two goals: first, to argue that transracial adoptees lack the critical resources to adequately articulate their experiences, which constitutes a hermeneutical injustice. Second, to point towards potential strategies or ways of thinking that could assist adoptees in navigating their experiences which are yet to be widely recognised, both individually and as a community. As a general rule, an individual is a transracial adoptee if they are an adoptee who is of a different race to at least one of their parents who are most likely part of the dominant racial and cultural group. Here, I am clearly casting the net fairly wide but this is intentional. Adoptees have vastly disparate relationships to their parents, their dominant culture, their ethnic, racial and cultural heritage and so on. Therefore it is a clearly hopeless task to try and articulate experiences that all adoptees have; this is inconceivable and moreover, unhelpful. The goal then is to articulate adoptee specific experiences that arise as a unique result of their positionality. What then, if anything, is common in the experiences of transracial adoptees? Adoptees are forced to draw on conceptual frameworks that only speak to part of their identity. As a result there are few conceptual resources that adoptees can draw on that capture their experience at this intersection. I will demonstrate how this results in systematic hermeneutical injustice. I will then pose the question: what sort of frameworks or tactics should we be employing to engineer and understand concepts that will help better concretise the transracial adoptee experience? I will then discuss the strengths and drawbacks of a decentered 'hometactic' approach which draws on work by Mariana Ortega. To ground the discussion I will be using Jessica Walton's concept of (Re)embodiment from her work on the experiences of South Korean adoptees. I will conclude by talking about the possibility of coalitional politics, both intra and inter-group alliances that adoptees would benefit from, but also point out some potential concerns about assimilation and reductionism.

2. Hermeneutical Injustice

Hermeneutic Injustice is a term coined by Miranda Fricker that aims to pick out a specific form of epistemic injustice. Fricker defines the term as "the injustice of having some significant area of one's social experience obscured from collective understanding owing to a structural identity prejudice in the collective hermeneutical resource"[126]. This is owing to the unequal distributions of power between groups when collective social meanings are generated and disseminated, the social positions of some groups leads to unequal hermeneutical participation[127]. When one group is subject to unequal hermeneutical participation with respect to "some significant area(s) of social experience, members of the disadvantaged group are hermeneutically marginalised"[128]. Consolidating these points, we can say that an individual experiences hermeneutical injustice when they:

1. Participate in a culture with social structures that lacks the appropriate concepts or hermeneutical resources to accurately portray an aspect of their experience.
2. Are actively harmed or disadvantaged by this lack of hermeneutical resources.
3. Are affected on a structural level, that is, that this gap in relevant hermeneutical resources is due to membership in a certain social group that is hermeneutically marginalised.[129]

[126] Miranda Fricker (2010) *Epistemic Injustice*, Oxford: Oxford University Press, 155.

[127] Fricker, 152.

[128] Fricker, 153.

[129] Thank you to an anonymous reviewer for suggesting that I defend this particular definition of hermeneutic injustice. In short, I do not believe that Fricker's account is *better* than the supplementary accounts of hermeneutic injustice given by other philosophers, such as Kristie Dotson or Ariana Falbo. This account could be strengthened by the concepts of contributory justice (Dotson, Kristie (2014) 'Conceptualising epistemic oppression', *Social Epistemology*, *28*(2), 115–138. https://doi.org/10.1080/02691728.2013.782585) and/or Positive/Negative Hermeneutic Injustice (Falbo, Arianna (2022) 'Hermeneutical injustice: Distortion and conceptual aptness', *Hypatia*, *37*(2), 343–363. https://doi.org/10.1017/hyp.2022.4). My use of Fricker here relates directly to a hermeneutical lacuna, rather than the denial of uptake (Dotson) or the construction of contradictory controlling images or oppressive distorting concepts (Falbo). Thus, it fits the purposes of my discussion. One potential avenue for thinking about this would be to analyse the distorted image of the adoptee in the public consciousness and media, such as in Hübinette, Tobias (2020) 'When the others other: Images and representations of transnational adoptees of colour among non-adopted Swedes of colour as reflected in contemporary Swedish minority literature', *Adoption &amp; Culture*, *8*(2), 245–264. https://doi.org/10.1353/ado.2020.0006

Let us quickly demonstrate with one of Fricker's examples. Carmita Wood worked an admin desk job in Cornell University's department of nuclear physics. One professor "seemed unable to keep his hands off her" and engaged in behaviour on multiple occasions that we would now term 'sexual harassment'[130] (Brownmiller in Fricker, 150). Wood was left stressed and traumatised and eventually resigned from her job but was denied unemployment insurance because she was unable to accurately describe her experience; "Wood was at a loss to describe the hateful episodes". Only much later, in a group with a number of other women who had all had similar experiences, did they finally give a name to this phenomena; "Somebody came up with 'harassment'. Sexual Harassment! Instantly we agreed. That's what it was"[131].

Wood was unable to articulate her experience without the concept of sexual harassment, which satisfies 1. She was denied unemployment insurance on this basis and was thereby harmed, fulfilling 2. And finally 3 is satisfied because the gap in hermeneutical resources Wood experienced was due to the hermeneutical marginalisation of women as a social group (many other women experienced similar wrongs and had no way of conceptualising it).

3. Constructing the Transracial Adoptee Standpoint

This section will aim to (provisionally) provide and answer two key areas of inquiry:

1. What is significant and/or unique to the transracial adoptee's positionality?
2. How does this create hermeneutical injustice? What are some examples of adoptee specific experiences that could be understood as 'properly our own'[132]?

Our initial task will be to roughly demarcate the boundaries of the group.

In preparation, we must begin by elucidating the adoptee positionality itself and its relationality to other social groups and standpoints. I am adopting the notion of relationality presented by Sarah Hoagland. For Hoagland, relationality is a key

[130] Fricker, 150.
[131] Fricker, 150.
[132] This is outside the scope of this paper but I hold that transracial adoption problematizes the very notion of proper ownership.

aspect of the way "our subjectivities are formed through our engagements with each other, both individually and culturally"[133]. That is, the social relations and social identities that construct knowers qua subjects do not exist separately and autonomously but are rather formed through their interactions with the Other. For example, races and racialised subjects do not pre-exist their relationality with other races: "Whiteness does not exist independently from engagements with people of colour"[134]. Crucial to Hoagland's account is that she is making a strong claim about the ontological status of racialized subjects as well as an epistemological claim about the relationship between knowers and objects of knowledge; "relationalities are rendered invisible through an epistemology that presupposes autonomy and denies relationality between knower and known"[135]. Social relations are, in fact, ontologically constitutive of social categories such as "white" and "colonised", but this is constantly ignored through an epistemology of ignorance; "That (most) whites walk through our day ignorant of our interdependency with peoples of colour is not about the invisibility of whiteness but rather about the erasure of peoples of colour as subjects"[136]. This ignorance of relationality, as we will see further on, is a serious impediment to adoptees insofar as they wish to know themselves.

Acknowledging this relationality is of paramount importance to understanding the social position of transracial adoptees. These adoptees are positioned at the intersection of white (predominantly) enculturation and their other racial and/or ethnic identity. As a result, the social location of the adoptee cannot be collapsed or neatly encapsulated by that of their white family members or their racial heritage. One way in which this can be illustrated further is by understanding the sphere in which each identity becomes salient: more often than not, the white encultured identity is salient in matters of private family life while the racial identity becomes more salient in public settings (walking around as the only minority in the family, having people treat you as a racial minority at school or in the workplace, acknowledged by other members of the racial group as being 'one of them' etc.). This

[133] Hoagland, Sarah Lucia (2007) 'Denying Relationality Epistemology and Ethics and Ignorance', in Shannon Sullivan and Nancy Tuana, eds., *In Race and Epistemologies of Ignorance*: 95–118. State University of New York Press, 97.

[134] Hoagland, 97.

[135] Hoagland, 97.

[136] Hoagland, 97.

is quite clearly a unique set of experiences that arise as a result of the specific social relationality of transracial adoptees. However, as Hoagland alluded to before, knowledge of this relationality has been ignored and effaced historically, which has resulted in phenomena such as the transracial adoption paradox. The transracial adoption paradox is a phenomenon coined by psychologist Richard Lee and refers to a set of contradictory experiences that arise due to the fact that "adoptees are racial/ethnic minorities in society, but they are perceived and treated by others, and sometimes themselves, as if they are members of the majority culture (i.e., racially White and ethnically European) due to adoption into a White family"[137]. There is therefore a serious tension in the ways in which transracial adoptees identify; "'I'm Australian but I'm not — I'm Korean but I'm not — I'm White but I'm not — I'm Asian but I'm not — I'm the same but I'm not"[138]. The claim that we are 'the same but not' points to an inherent fragmentation in the adoptee identity[139]. This is a problem because it then forces us to draw from conceptual resources from either racial pools, neither of which can accurately articulate this set of experiences.

Keeping in line with our areas of inquiry I will now quickly demonstrate how this constitutes a hermeneutical injustice (although I believe the immediate intuition is quite strong). As demonstrated above, transracial adoptees participate in a culture that lacks the appropriate hermeneutical resources to accurately portray their experiences. This is exacerbated by their unique relationality of being ignored or erased. In practice, this is a function of the adoptee's relatively subordinate status in relation to the rest of their family; many parents adopt a 'colour-blind' attitude that actively aims at erasing racial difference. Transracial adoptees do not have access to "'the communal nature of racial melancholia' precisely because there is no 'intergenerational and intersubjective process' of recognizing and affirming that experience"[140]. The adoptee is therefore hermeneutically marginalised by being denied access to the relevant cultural and racial communities that would be able to

[137] Lee, Richard M (2003) 'The Transracial Adoption Paradox: History, Research, and Counselling Implications of Cultural Socialization', *The Counseling Psychologist 31, no 6*: 711.

[138] Heaser, E. HeeRa (2016) *Korean Australian Adoptee Diasporas: A Glimpse into Social Media.* [Doctoral dissertation, University of New South Wales], 194.

[139] As constructed using dominant understandings of identity (either being white *or* otherwise, for example). I would argue that this fragmentation is due to the forgetting of relationality.

[140] Gustafsson, Ryan (2020) 'Theorising Korean transracial adoptee experiences: Ambiguity, substitutability, and racial embodiment', *International Journal of Cultural Studies*, *24*(2), 316.

perhaps help in articulating these issues. Adoptees experience the existential ennui of grappling with a fragmented and ambiguous identity that does not fit because their unique relationality is ignored: many adoptees "fe[el] trapped by the expectation to choose to be either Korean/Asian or Australian/White…I don't really want to associate with any of them [Korean or Australian]. I just want to be myself"[141]. This effacement in the family setting also leads to tangible material and psychological harms. Without the relevant support from their family in trying to understand their racial identity, many adoptees experience worse levels of well-being: In a psychological study of 34 transracially adopted Korean American youths living with White parents in the United States, Diane Lee found that "the resilience that [Korean Adoptees] display is intricately tied and perhaps dependent on the support that they receive from their White parents…it was found that two specific aspects of family warmth, cohesion and conflict, are most important in fostering the psychological resilience and flourishing of transracially adopted Korean youths"[142]. Therefore, this lacuna in hermeneutical resources is a form of structural hermeneutical injustice experienced by transracial adoptees.

At this point I want to clearly lay out the rationality behind my examples moving forward. This section is predominantly exegetical, Ryan Gustafsson's importance here cannot be overstated. I will outline two concepts put forward by Gustafsson: Hyper(in)visibility and Epistemic Ambiguity. These will be discussed relatively briefly for the following reasons: first and foremost, I wish to include them because I think that they add an interesting and rich layer of understanding to the situated experiences of transracial adoptees outlined above. Additionally, I aim to open these concepts up to potential future analysis; these examples will hopefully open up questions to which I do not have the answers to and/or fall outside of the scope of this paper. They represent an important contribution to this new area of work — a philosophical understanding of transracial adoptees. The next section will then begin with a discussion of another concept, (Re)embodiment, which I believe better fits within the bounds of the current paper which I will use as a case study for

[141] Heaser, 'Korean Australian Adoptee Diasporas', 195.

[142] Lee, Diane Sookyoung (2016) 'The Resilience of Transracial Korean American adoptees: Cultural identity crisis within the family and the mediating effects of family conflict and cohesiveness during adversity', *Adoption Quarterly*, *19*(3), 161.

evaluating Mariana Ortega's notion of 'hometatctics' as a way of solving the problem.

The notion of Hyper(in)visibility is put forth by Ryan Gustafsson in their work in order to "capture this sense of simultaneous exposure and hiddenness, but also to emphasise how, for transracial adoptees, visibility is achieved via invisibility"[143]. In tandem with the erasure of racial identity and the forgetting of relationality, adoptees are subject to a paradoxical phenomenological schema of visibility. The adoptee is both highly visible and also presented as totally invisible at the same time, across different social contexts. For example, the transracial adoptee is hyper-visible within their family, standing out as a different race, but as aforementioned, this difference is sometimes denied and unacknowledged, thereby rendering the adoptee invisible. From this, Gustaffson concludes that "the adoptee's visibility is achieved through, or at the price of, invisibility and vice versa…in order to be visible, the adoptee is made to disappear, or is made invisible"[144]. This is due to the particular position of the adoptee; the dissonance between outward racial presentation and internal white enculturation or rather, the understanding that they have not been raised in a racialised environment and are therefore not privy to certain knowledge (of customs, language, food etc). As Gustafsson describes it, this means that the adoptee body in her home country is visible as racially different, which necessarily means that their adoptee identity and their white enculturation becomes invisible. An identical logic is at work in the opposite case: a Korean adoptee in South Korea might 'blend in with the crowd' and as a result their racial difference is made invisible. However, the adoptee knows that they do not fit in seamlessly and this is "often accompanied by experiences of 'standing out' via other signs, most commonly, by one's inability to speak Korean fluently"[145]. The becoming-visible of one aspect of the adoptee identity is made possible only by the making invisible of the other aspect: at home, the body is marked for difference while in their country of origin, blending in "render[s] visible one's difference (to oneself), reinforcing and

[143] Gustafsson,'Theorising adoptee experiences", 318.
[144]Gustafsson, 'Theorising adoptee experiences' 318.
[145] Gustafsson, 'Theorising adoptee experiences", 319.

amplifying it[146]. In both cases, visibility and invisibility function to make hyper-visible one's difference (to oneself) as an adoptee."[147]

Epistemological Ambiguity characterises how transracial adoptees in the actual world are denied stable epistemological footing when interrogating matters of their origins. Adoptees are a product of social institutions, rituals and legal codes which by its very form, entails a disconnect of the adoptee from their racial, ethnic, and genealogical origins; "In order to facilitate the potential adoption of a child, adoption agencies had to in effect create orphanhood administratively…This becoming-bare of the child, which is also the becoming-adoptable of the child, hence entails a legal and social severance or detachment from natal parents, siblings, and relatives"[148]. The actual institutionalisation of adoption and the intervention of adoption agencies means that in search of answers, the adoptee constantly has the epistemic floor dragged out from under their feet; "it is important to note that the institutional processes mentioned form part of the historical and social context within which adoptees attempt to forge 'knowable' or legible individual life-histories"[149]. Without reliable ways of cross-checking and verifying information given by the agency, adoptees seeking information are stranded and forced to operate within a terrain of uncertainty; "Epistemological ambiguity stems from not just the absence of knowledge but also the impossibility of knowing and, relatedly, the ambiguous value or status of any knowledge gained"[150]. In a more general sense, this

[146] Gustafsson, 'Theorising adoptee experiences', 319.

[147] A question one may have at this point is how the experience of transracial adoptees differs from an experience of white passing. While they both operate within a logic and discourse of racial visibility, I believe there are very clear differences. The main difference is the relationship between racial appearance on the surface of the body and that body's 'truth' in a sense. That is, when a light-skinned black person who passes as white is called out to: ""Hey, white girl! Give me a quarter!"(Piper, Adrian (1992) 'Passing for white, passing for black', *Transition*, (58), 4–32. https://doi.org/10.2307/2934966), there is a sense in which the caller is getting something *wrong*. Adrian Piper is *not* white, this person has made a mistake. On the other hand, when the auntie at my local Korean restaurant speaks to me in Korean, expecting me to understand, she is *not* getting something wrong in the same way. My racial presentation is simultaneous with the 'truth', I am Korean. There is no sense of passing here in the traditional sense. If there is a feeling of 'passing as Asian' for example for transracial adoptees then it will arise from a lack of cultural knowledge, not necessarily from being mis-recognised as being a race which they are not. There may be space to further interrogate the relationship between the two phenomena but that would fall outside the scope of this paper. At this point they are two very distinct phenomena that both operate within a similar discourse of race, visibility etc.

[148] Gustafsson, 'Theorising adoptee experiences', 312.

[149] Gustafsson, 'Theorising adoptee experiences', 312.

[150] Gustafsson, 'Theorising adoptee experiences', 312-313.

epistemological ambiguity is prominent in the seemingly innocuous questions that adoptees get asked everyday but do not have immediate answers: "who are your real parents?", for example. This poses similar questions about genealogical origins, who are my ancestors? Where or what is my history? Answers to these questions are not obvious and fall outside of the aims of this paper but I believe it to be a really insightful aspect of adoptee identities.

4. Models Going Forward

In this section I will be looking at Mariana Ortega's notion of 'hometactics' as a model for overcoming this hermeneutical injustice. To ground the discussion I will be drawing on the concept of (Re)embodiment.

(Re)embodiment is a term put forward by Jessica Walton in her anthropological study of Korean adoptees who have chosen to return back to Korea. Walton argues that many Korean adoptees move back to Korea as a way of better understanding their Korean identity; "Korean adoptees attempt to (re)embody their physical bodies by trying to identify with their appearance as well as their past. These acts toward (re)embodiment demonstrate agency as Korean adoptees try to make a Korean identity something that they can feel, move around in, experiment with and understand"[151]. By actually living and being in Korea, eating the food and participating in the culture, adoptees are able to better understand the facts of their Korean origins through lived experience[152]. Borrowing from Thomas Csordas, Walton understands embodiment as "an indeterminate methodological field defined by perceptual experience and mode of presence and engagement in the world"[153]. Walton claims that adoptees are able to re-embody their Korean 'mirror image' by actually travelling to, and living in Korea, which makes this identity more real and tangible. By dressing a certain way, eating certain foods and adopting certain mannerisms, adoptees are employing tactics that "make a Korean identity a part of their sense of self. Eating Korean food is a way to embody a Korean Identity"[154]).

[151] Walton, Jessica Rose SeeYoung (2009) *(RE)EMBODYING IDENTITY: Understanding Belonging, 'Difference' and Transnational Adoption through the Lived Experiences of Korean Adoptees.* [Doctoral dissertation, University of Newcastle], 250.

[152] Walton, '(RE)EMBODYING', 255.

[153] Csordas in Walton, '(RE)EMBODYING', 255.

[154] Walton, '(RE)EMBODYING', 269.

Moreover, it is necessary to recognise that not having knowledge of these things (how to dress, what to eat, etc), produces a profound feeling of not being Korean, of being a fake. Consequently,acquiring this knowledge and applying it in practice allows a Korean identity to emerge in praxis, that is, "Korean identity is not given, but something that has to be worked through…By being in Korea and literally embodying a Korean identity through food, they are trying to feel a meaningful Korean identity through their body"[155].

5. Hometactics

Mariana Ortega's work is based in a politics of location and a phenomenology of the home. For Ortega, the question of what it means to belong, to dwell in the world and to have a home which is connected to an 'authentic belonging' is problematised by what she calls the multiplicitous self[156]. The multiplicitous self is the self that occupies multiple positionalities in terms of gender, race, class and so on simultaneously and is therefore capable of "occupying a liminal space of space of in-betweenness"[157]. The question of home for the multiplicitous self then becomes a question of homes. The notion of 'hometactics' is aimed at shedding light on the actual praxis of everyday life, the 'making-do' that multiplicitous selves undergo all the time in order to "negotiat[e] their multiple identities in light of both ambiguities and contradictions"[158]. Hometactics emphasise the practical aspect of making the world 'homely' for those without a stable home, it draws attention to ways of being in the world, modes of living with the ambiguities and the contradictions. Although this may entail that multiplicitous selves do not form stable and robust senses of belonging to a single 'home', this should not undermine the ability to create meaningful social and political bonds and coalitions of resistance; "The sense of individual or group 'belonging' that they may provide is a great source of comfort in the midst of the complex, sometimes ambiguous, sometimes contradictory lives of multiplicitous selves"[159].

[155] Walton, '(RE)EMBODYING', 279-280.

[156] Ortega, Mariana (2014) 'Hometactics', in Emily Lee, ed, *Living Alterities*: 173–88.

[157] Ortega, 'Hometactics', 176.

[158] Ortega, 'Hometactics', 181.

[159] Ortega, 'Hometactics', 185.

For me, it is clear that practices of (Re)embodiment are an archetypal hometactic. (Re)embodiment details quite literally the experiences of adoptees 'making-do' with what they have and their legitimate attempts to make sense of their identity. Ortega's 'multiplicitous self' is also useful here in helping us navigate these complex and ambiguous situations that we have to confront all the time. With reference to our discussion of Epistemic Ambiguity above, the epistemically shaky grounds upon which questions of origin can be understood clearly problematises the question of home: where is home for the transracial adoptee? Common questions such as 'who are your real parents?' only exacerbate this confusion. Framing how adoptees navigate their racial identity through the notion of 'hometactics' accentuates its significance; the desire to engage in the culture and embody a racial identity is not simply a curiosity nor is it a simple way of understanding oneself better. Rather, it is a crucial aspect of one's ability to dwell within the world, to survive, to flourish, to create and to resist.

6. Coalitional Politics

Here is a potential problem for Ortega: it seems to be a short term solution. That is, while it has useful explanatory power regarding the home-making practices of 'making-do', it seemingly has little normative force which directs us towards what we ought to do, especially in the context of a long-term political project. While it is all good and well to describe these practices, the model itself does not seem to be able to critique these practices or revise concepts if needed. Although they can be done collectively, hometactics are a largely individual task, it is about making do with what I have and creating a sense of belonging for myself. If this is threatened or criticised in any way, this could be seen as a direct attack on one's sense of place and belonging in the world; in this scenario, hometactics would be a hindrance to new thought. As Ortega herself recognises, hometactics "do not form a robust sense of belonging or familiarity, whether it is associated with a location or a group, and thus they might not be capable of forging strong political coalitions that can establish practices of resistance"[160].

[160] Ortega, Mariana (2016) *In-Between*, SUNY Press, 205-206.

In terms of relieving hermeneutical injustice, hometactics can help with clarifying our current practices and ways of engaging with the world which allows us to bridge the hermeneutical gap. It is clear that hometactics are a good way of framing (re)embodiment practices such that we are better at articulating our estranged sense of self. Our identities are heterogenous and ambiguous hence we struggle to find a place where we belong, a home. Nonetheless, I would maintain that this is predominantly due to the insufficient conceptualising of the adoptee experience and the forgetting of relationality. This is an integral line of thought in this paper for me: although the vast majority of the experiences I have outlined have been ones of negativity, tension and being out-of-place, I am not a pessimist about the future. Rather, I believe that with further thinking, the goal will be to eventually dispense with these concepts as expressions of the transracial adoptee phenomenology because we have formulated a more robust sense of identity and solidarity such that issues of being 'not Korean' or 'not Australian' enough, for example, are less significant. At this juncture then, I wish to speak of coalitional politics.

Here I will look at Ortega's understanding of coalitional politics and an ally in Maria Lugones and her notion of deep coalition. But first one may think that I need to justify the need for broader group politics instead of the cultivation of our own small community. My answer would be that we do not live in an adoptee-only bubble nor should we aspire to; to take that as an ideal would be to reaffirm a politics of liminality, of a self-contained space. But this is not the world we live in — we must learn to live together with others. Not only that, but many of our struggles are predicated on our marginalisation from groups that we actually do belong to, most prominently that of our racial/ethnic identity. Thus I wish to briefly outline some considerations regarding coalition between, say, between Asian-Australian adoptees and non-adopted Asian-Australians and what sorts of understandings and channels of communication should be open for our marginalisation to be recognised and addressed.

Ortega identifies some key elements of coalitional politics: an understanding that coalitional politics is about both being/belonging and becoming "which includes location, being-with, and becoming-with", as well as a recognition of both shared oppression and resistant agency, which is dependent on what Lugones theories as "complex communication" that can lead to "deep coalition"[161]. As Ortega makes clear, we must be mindful "of the manner in which groups are already heterogenous" so that we do not fall into the plight of simple identity politics which assigns group identity based on homogeneity; rather, we should take heed of the call for "basing identity on politics rather than politics on identity". For Lugones, complex communication begins with a mutual recognition of liminality but the simultaneous recognition of difference — we do not understand each other transparently in virtue of both occupying liminal subjectivities. Thus, complex communication with the other requires the mutual recognition of a marginalised position but also requires that we be "disposed to understand the different ways in which others communicate and resist without trying to assimilate or reduce them to our language and to ourselves"[162]. Jose Medina aims to combat this problem with what he calls the 'kaleidoscopic consciousness'. Rather than a double-consciousness or an infinitely pluralised consciousness, Medina advocates for a consciousness "that has built into it a flexible and dynamic structure so that it can always adapt to the possibility of excess, that is, of there being more ways of experiencing the world than those considered"[163]. The kaleidoscopic consciousness believes that "ready-made meanings and fixed frameworks of intelligibility fail us"[164] which, for our purposes, will be helpful in trying to show to non-adopted members of the racial group that adoptee experiences are racialized experiences in a legitimate way i.e, Asian adoptee experiences are still Asian experiences!

Here it is important to emphasise a meta-point about the thesis as a project: I have tried to distinguish certain aspects of the adoptee experience as irreducibly our own, that is it can be very tempting for second-generation immigrants or mixed-race people to claim or identify with adoptee experiences. I think that this is both insightful and helpful, however I would like to maintain that there are aspects of our experience that are unable to be assimilated into the experiences of others in a

[161] Ortega, *In-Between*, 163.
[162] Ortega, *In-Between*, 166.
[163] Medina, José (2020) 'Complex Communication and Decolonial Struggles: The Forging of Deep Coalitions through Emotional Echoing and Resistant Imaginations', *Critical Philosophy of Race 8*, no. 1–2: 200-201.
[164] Medina, 'Complex Communication and Decolonial Struggles', 200.

one-to-one way. This is what Medina calls 'blindness to differences'[165]. In short, there is the worry that members of the relevant racial groups will not realise that there are important intra-group differences that can be covered with the blanket term 'Asian' for example. But on the other hand, there is also the worry that adoptee experiences will be entirely overlooked as 'inauthentic' experiences due to white enculturation. It is here then, that I am speaking not to adoptees but instead to non-adopted people of the relevant racial/ethnic groups. Asian adoptees are Asian! Accordingly, I would call for the necessary re-examination of 'standpoints' to accommodate adoptees, deep coalition through complex communication seems to me to be an important development for racial groups which will encompass a wider range of experiences and allow for new ways of becoming-with; " There is no complex communication if the communicators come out of the encounter untouched, with their subjectivity unaltered"[166]. These groups have a lot to teach each other[167].

One criticism that could be made is that this picture paints the adoptee subject as being too passive, as too dependent on the recognition and action of others. In brief, I can almost bite the bullet on this criticism. I believe that a key proactive move on the part of the adoptee is embodied in the notion of 'hometactics' so this criticism mainly applies to our place in coalitional politics. I believe that it is reasonable to assume that adoptees cannot do all of the work themselves, especially when we recall the importance of relationality — we are constructed in conjunction with others. We might then be able to call for a model of shared responsibility (also following Medina) but this again, would fall outside the scope of the thesis.

7. Conclusion

The core aim of this paper was to justify the classification of the conceptual lacuna regarding transracial adoptees as a hermeneutic injustice. I outlined what I thought were the main issues, namely what made the transracial adoptee standpoint distinctive and what sort of experiences we could draw as a result. I argued that the adoptee standpoint is distinguished by its unique social positionality and its

[165] Medina, 'Complex Communication and Decolonial Struggles', 208.
[166] Medina, 'Complex Communication and Decolonial Struggles' 212–36.
[167] Going any deeper into what this might look like or issues of epistemic appropriation fall outside of this thesis.

relational nature, situated at the crossroads of white enculturation and racial minority identity. I looked at a 'hometactic' model which may help to guide future inquiry. I then raised the concern that hometactics may struggle as a long-term communal political project and I therefore introduced Maria Lugones' notions of coalitional politics and complex communication. The experiences of transracial adoptees may be brushed aside in multiple ways; non-adopted members of the relevant racial/ethnic group may reject the perceived 'authenticity' of the experience or, on the other hand, there may be a temptation to assimilate or reduce the adoptee experience to that of their own. This is terrain that is vastly unmapped and very new. My intention was to begin to sketch its silhouette, erase and redraw some misguided lines and paint just a small section of the landscape.

References

Dotson, Kristie (2014) 'Conceptualising epistemic oppression', *Social Epistemology*, *28*(2), 115–138. https://doi.org/10.1080/02691728.2013.782585

Falbo, Arianna (2022) 'Hermeneutical injustice: Distortion and conceptual aptness', *Hypatia*, *37*(2), 343–363. https://doi.org/10.1017/hyp.2022.4

Fricker, Miranda (2010) *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Gustafsson, Ryan (2020) 'Theorising Korean transracial adoptee experiences: Ambiguity, substitutability, and racial embodiment', *International Journal of Cultural Studies*, *24*(2), 309–324. https://doi.org/10.1177/1367877920938374

Heaser, E. HeeRa (2016) *Korean Australian Adoptee Diasporas: A Glimpse into Social Media.* [Doctoral dissertation, University of New South Wales].

Hoagland, Sarah Lucia (2007) 'Denying Relationality Epistemology and Ethics and Ignorance', in Shannon Sullivan and Nancy Tuana, eds., *In Race and Epistemologies of Ignorance*: 95–118. State University of New York Press.

Hübinette, Tobias (2020) 'When the others other: Images and representations of transnational adoptees of colour among non-adopted Swedes of colour as reflected in contemporary Swedish minority literature', *Adoption &amp; Culture*, *8*(2), 245–264. https://doi.org/10.1353/ado.2020.0006

Lee, Diane Sookyoung (2016) 'The Resilience of Transracial Korean American adoptees: Cultural identity crisis within the family and the mediating effects of family conflict and cohesiveness during adversity', *Adoption Quarterly*, *19*(3), 145–165. https://doi.org/10.1080/10926755.2015.1121184

Lee, Richard M (2003) 'The Transracial Adoption Paradox: History, Research, and Counselling Implications of Cultural Socialization', *The Counseling Psychologist 31, no 6*: 711–44. https://doi.org/10.1177/0011000003258087

Lugones, Maria (2003) *Pilgrimages = peregrinajes: Theorising coalition against multiple oppressions*. Rowman & Littlefield Publishers, Inc.

Medina, José (2013) 'The epistemology of resistance: Gender and racial oppression, epistemic injustice, and resistant imaginations', *Choice Reviews Online*, *50*(11). https://doi.org/10.5860/choice.50-6107

Medina, José (2020) 'Complex Communication and Decolonial Struggles: The Forging of Deep Coalitions through Emotional Echoing and Resistant Imaginations', *Critical Philosophy of Race 8*, no. 1–2: 212–36. https://doi.org/10.5325/critphilrace.8.1-2.0212.

Ortega, Mariana (2014) 'Hometactics', in Emily Lee, ed, *Living Alterities*: 173–88

Ortega, Mariana (2016) *In-Between*. SUNY Press.

Piper, Adrian (1992) 'Passing for white, passing for black', *Transition*, (58), 4–32. https://doi.org/10.2307/2934966

Walton, Jessica Rose SeeYoung (2009) *(RE)EMBODYING IDENTITY: Understanding Belonging, 'Difference' and Transnational Adoption through the Lived Experiences of Korean Adoptees.* [Doctoral dissertation, University of Newcastle].

Literature as a Pre-Philosophy: Exploring Julian Marias's Notion of Dramatismo and Narrative

Francisco Silayan Pantaleon[168]

University of Asia and the Pacific

Abstract

Spanish philosopher Julián Marías explains that the adequate philosophical explanations of the human person reside in literature, particularly in the constitutive *dramatismo* (dramatic character) of the person, which is made meaningful by narrating human life. He claims that literature is a sort of pre-philosophy, as has been the case since the time of the Greeks, especially in their presentation of philosophy in the form of literature, that is, the story-like structure of the dialogues. Marías says life has *dramatismo* because it consists of a series of circumstantial happenings that have a projective quality, and this is only intelligible through narration, by 'giving an account' of the dramatic character of my life. Since my life is a story on account of its *dramatismo*, it is only properly recounted, that is, understood, when it is narrated. But no matter how much these two literary notions inform philosophical inquiry, they can never be isolated from their proper domain: literature. In some way, then, philosophy relies on literature because of the ease with which it penetrates the reality of the human person; and the tools that make it possible are, as I shall explore in this paper, Marías's notions of *dramatismo* and narrative.

[168] Francisco Silayan Pantaleon is taking his Master's in Humanities at the University of Asia and the Pacific (UA&P), Philippines. He's a teaching assistant for the Department of Philosophy of UA&P and is completing his thesis on the philosophy of Julián Marías as justification for the study of the humanities. His teaching and research interests include personalist philosophy, metaphysical anthropology, and the history of the humanities and the liberal arts.

1. Introduction: The Adequate Concepts from a Pre-Philosophy

Following his mentor José Ortega y Gasset, Spanish philosopher Julián Marías attributes to the human person a circumstantial character, such that the reality of the human person cannot be understood apart from his circumstance, nor can his circumstance be understood apart from him. The reason is that "a self can never be postulated as an ontologically independent being".[169] Hence, Marías, in the course of developing his own philosophy, adopted the formula developed by Ortega to describe—without presuming to exhaust—the reality of the human person: *Yo soy yo y mi circunstancia*, that is, 'I am I and my circumstance'.[170] This is why it is possible to understand the human person as someone who "acquires the ultimate circumstantial and individual reality, the absolutely concrete reality, of *each* life, which *happens* dramatically, in respect to which the possible and adequate form of 'enunciation' is to *narrate it*".[171] Two ideas are of great importance here, which will prove to be the starting points for understanding how literature is a precursor to philosophy. First, that human life, because it is circumstantial, happens dramatically; second, that the adequate method of speaking of life's dramatic quality is to narrate it. From these two key aspects by which the reality of the human person is made manifest, we find Marías referring to two literary concepts to discover the person: *drama* and *narrative*.

Such a curious deference to literary concepts, which will turn out to offer fantastic philosophical nuances, is deliberate on Marías's part. For, in one of his works, he makes the bold claim that literature, particularly the novel, is a 'pre-philosophical' method by which we can access the reality of the human person.[172] But in what way is literature prior to philosophy as a 'pre-philosophy'? "Do not forget", Marías writes, "that the intellectual, philosophical discovery of human life has been *posterior* to the creation of a splendid literature [my emphasis]".[173] He offers, as examples, the Homeric poems, the stories of the Bible, the Bhagavad Gita, the

[169] Mora, José Ferrater (2003) *Three Spanish Philosophers: Unamuno, Ortega, Ferrater Mora*, State University of New York Press, 147.
[170] Marías, Julián (1971) *Metaphysical Anthropology: The Empirical Structure of Human Life*, Frances López-Morillas, trans, The Pennsylvania State University Press, 71.
[171] Marías, *Metaphysical Anthropology*, 76.
[172] Marías, Julián (1996) *Persona*, Alianza Editorial, 65, this and succeeding quotes from *Persona* were translated by Paul Dumol (August 2023).
[173] Marías, *Persona*, 82, my emphasis.

Qur'an, and countless historical narratives. In other words, he means to say that literature is prior to philosophy in the discovery of the human person, that is, literature was talking about human persons long before philosophy began to. While philosophy began with concepts proper to things (e.g., Aristotle's *ousía*, the Scholastic's substance, Descartes's *res cogitans*, and Heidegger's *Dasein*), literature began with stories and myths about human persons (e.g., Homer's *Iliad* and *Odyssey*, Virgil's *Aeneid*, Dante's *Divine Comedy*, etc.).

Literature has always assumed its chief subject matter to be about *persons*, about human life.[174] Hence, it is possible, through literature, to reach into the person and speak of him, make an accounting of him. But this is achieved only by having recourse to the adequate concepts that unveil this marvelous reality, which, for Marías, involves *drama* and *narrative*. Marías does not give explanations as to why drama and narrative are among the 'adequate concepts' by which the reality of the human person is made manifest, but it seems to me the proof lies precisely in the richness that is drawn from the reality of the human person when these concepts are used to understand him: that human life, because it is dynamic and ongoing, has a dramatic structure that must give an account of itself, that must be narrated.

To elucidate these two concepts, we will have to lay out Marías's notion of *dramatismo* (dramatic character) and narrative, defining each one, describing the various aspects that constitute them, illustrating how they unfold in human life, in each one's life, and therefore serve as ample justifications for literature being a precursor to philosophy.

2. Dramatismo: A Constitutive Property of Human Life

The Spanish word *dramatismo*, which Marías uses, has no equivalent English translation, and the best approximation is 'dramatic quality' or 'dramatic character'. This *dramatismo* of human life is consistent with one of the first facts we come to terms with: that we do things and things happen to us. (Which is, uncoincidentally, Marías and Ortega's definition of *life*: "What I do and what

[174] "History and literature have taken as their great assumption their being about persons": Marías, *Persona*, 82.

happens to me").[175] From these 'happenings' in my life I discover that these things that 'surround' me are the concrete materials in my life that make up the 'stage', 'setting', or 'mise-en-scène' of my life. This Marías calls the *circumstances* where I find myself in life because each one's "circumstance… is not a collection of things, but a stage or world where this drama [of each one's life] is played."[176] Each one finds himself in the drama of his own life, an exclusively *personal* drama that projects to the future with a plot-like structure.

My life moves forward as in a plot because my circumstances change. We are speaking here of the projective quality of human life, which indicates that human life begins some*where* (my 'stage' or circumstance) and moves forward, by making use of the things I find in my circumstance, to 'somewhere else' (my next stage made up of a new circumstance thanks to the movement of my life, that is, my *living*). This projective quality, life's 'ongoingness' is suggestive of the future-oriented characteristic of human life, so much so that I cannot begin to think of my 'where' *now* without a view of the 'where' *I plan to be*.[177] In other words, I cannot think of the present without thinking of the future, nor can I think of the future without thinking of the present. This is another way that makes the projective quality or the future-orientedness of human life unmistakably *dramatic*. Since my life is a project and living is projective, something I have to do *now* for the sake of *what will be*, I always have to deal with dramatic tensions in my life, which often have a character of uncertainty.

This uncertainty amplifies the *dramatismo* of my life because, through that very uncertainty (the precariousness of *what will be*), I must 'anticipate'. I must 'get ready' or 'fix my stance' on the stage that is my circumstances to encounter the next moment with some degree of preparation.[178] Now we are starting to see how drama is present in the project of human life, the anticipation that accompanies the progression of that project, and the projective quality of living that extends to the future with uncertainty. In all points of my life, therefore, there is a drama that

[175] Marías, Julián (1956) *Reason and Life: The Introduction to Philosophy*, Kenneth Reid & Edward Sarmiento, trans, Hollis & Carter, 207.
[176] Marías, *Metaphysical Anthropology*, 49.
[177] Marías, *Reason and Life*, 27-8.
[178] Cf. Marías, *Metaphysical Anthropology*, 243.

plays out—a drama in which *I* am the protagonist, because things happen to me, and I encounter the need to do something with things, to use the things that are present to me in my circumstance. It is thus fitting to say that *living* is inseparable from drama; that human life is inexplicable, unintelligible, without its inherent *dramatismo*.

We have expressed and explored a feature of *dramatismo* demanding that I do something with things in the ambit where I find myself. Its implications are illumined by the fact that my life 'loses', as it were, its dramatic character, or at least 'diminishes', when I do not do something with and make use of the things I find in my circumstance, and avoid or prevent things from happening to me. This 'interruption' in living, characterized by 'not doing anything', is the vital manifestation of the phenomenon we call 'boredom' and, as a consequence of boredom, 'idleness'. I first fall into boredom when I have nothing to do now, much more when I have no one to do things with.[179] It is as if I put living to a grinding halt. Moreover, boredom indicates an imminent (but temporary) biographical cessation—a 'postponement' of *living*—that soon ends up in a worse and more lasting state, idleness. Boredom and idleness are, therefore, the very antithesis of drama, the very absence of drama.

Further, we also observe that, in a literary sense — primarily in novels, theater, and films — there can never be room for boredom and idleness. Otherwise, the novel, play, or film's story ceases to be a story. There is no novel or film where drama ceases to be present, no kind of literature where boredom or idleness is present, because "all that is human can be quiescent, but never static".[180] There is always a happening, a doing, no matter how dull or uninteresting it may be. At *all moments* there is a drama ongoing because life is ongoing. Drama gives vitality to the story and is the principle of the story's movement and development. Perhaps this is one of the reasons why Marías posits that novels and films are most representative of human life.

[179] Cf. Marías, Julián (1993) *Razón de la filosofía*, Alianza Editorial, 123.
[180] Marías, *Metaphysical Anthropology*, 83.

We find in novels and films the very dramatic quality of living expected of each human life, and it is from them that I can learn to live up to my *dramatismo* and live dramatically so that I may be a dramatic event that is *really* 'living'. In novels and films, we discover a *story*; so, too, in human life, we discover a story because (1) in it there are elements of a story: setting, characters, moment, future, projection, and uncertainty, among others, and, more importantly, (2) human life is itself inherently and intimately *dramatic*, so much so that its *dramatismo*, although strictly speaking is a *characteristic* of human life, can be said, without any pretense of absolutization, as commutative to human life itself, i.e., that human life is drama and drama is human life.[181] Drama is an exclusively human feature: only persons can possess *dramatismo*.

Marías does not settle for a definition of "drama" but rather points out how it determines the structure of human life. Primarily, the implications on the dynamic character of human life that is always 'ongoing', or as Marías prefers to call it, 'arriving'. This should not be a cause for dissatisfaction because one can easily draw the conclusion from Marías that it was not his intention to offer a definition of *drama*. His conception of 'drama' is, as we have shown, *story-like*. It is not meant to be articulated with symbolic meanings enclosed in a genus with a specific difference, as 'definitions' are understood logically.

Rather, he sees the definition of drama as, itself, a story that unfolds, because, as he comments, "the myth [a story or a drama] is not something to fall back on in the absence of a definition, but something superior, in which genuine philosophical knowledge consists". And the reason for this is that a story is "something like an abbreviation [of knowledge about human life] accessible to man".[182] This 'lack of definition', so to speak, then becomes the very ingredient that helps us make fuller sense of what we have said about drama: that human life is drama and drama is human life. No other created beings are constitutively organized by a dramatic character apart from human persons.

[181] A simple syntactical observation of the following quotation will reveal the conclusion just made: "The only thing that interests human beings is human living, 'drama', and when this is lacking the film becomes a documentary and, whatever its virtues, produces boredom": Marías, *Reason and Life*, 64.

[182] Marías, Julián (1971) *Philosophy as Dramatic Theory*, James Parsons, trans, The Pennsylvania State University Press, 43-4.

3. Narrative: Accessing Human Life

My life, having a constitutive dramatic character, "is something that happens to me, here and now, in these precise circumstances, and the means of having access to it is to relate it, to tell someone about it. The form of 'statement' that corresponds to it is a report, a *narration*" [my emphasis].[183] This directs us to a more salient question worth attention and curiosity: what does *narration* consist of? In its commonest conception, it is something 'told' or 'said'. But narration includes an articulation of the *why* and the *how*—which, by their very semantic construction, have a temporal reference—and not simply an expression of *what*, which is a static matter-of-fact. It is one thing to say, 'I want a cookie', and another to say, 'I want a cookie *because* I am hungry'. The former is a plain utterance that evokes no hint of a story, of a drama, of a plot; but the latter, by its mere expression of the *why*—expressed in the reason: '*because* I am hungry'—is already a drama, or at the very least, has the proper ingredients for a drama.[184]

The reason for this is that, as Marías put it, "the narrative, the story, is the life-giving nucleus of the myth [or the drama]."[185] Drama is vivified—it 'comes to life'—when it is narrated. The drama of my life is concretized, incarnated, even in some way immortalized, when it is narrated, much more when *I* narrate it, that is, when *I* 'give an account' ('*dar razón*', in Spanish; literally, 'give reason') of *my* life. My life, as a story, re-counts the past from the present toward what remains of the future. This temporal distension of my life from my earliest recollections to what I anticipate is, precisely, the drama of my life concretely lived. Not anyone else's.

Hence, narration operates in a unitary fashion. A narrative speaks only of *one* life: this or that. When I narrate my life, for instance, I relate what *I* did with the things with which I found myself and why *I* did those. What we are saying at present will receive more clarity from the example Marías offers:

[183] Marías, Julián (1967) 'The Idea of Metaphysics', in Aloysius Robert Caponigri, ed & trans, *Spanish Philosophy: An Anthology*, University of Notre Dame Press, 363, my emphasis.

[184] "This structure could be formulated by saying that the past and the future are *present* in my life, in the 'why' and the 'wherefore' of each of my actions. In my immediate actions the past is present, because the reason for what I do can only be found in what I have done, and the future is present in the project, on which hangs the whole meaning of my life": Marías, Julián (1954) *Ensayos de teoría*, Editorial Barna, 48, my translation.

[185] Marías, *Philosophy as Dramatic Theory*, 44.

> […] I have compared the dictionary entries of three very different realities: for example, “pentagon,” “owl,” and “Cervantes.” Of the pentagon, an ideal object, the dictionary gives a *definition*; of the owl, a real object, a thing in the usual sense of the word, it gives a *description*; of Cervantes, a personal reality, it tells a *story*. The dictionary gives the “essence” of the pentagon: a polygon with five sides; it tells what the owl is, what it looks like, what it does, how it behaves—“the” owl, be it understood, “each” owl; but when it speaks of Cervantes it offers us a narration; it tells us where and when he was born, where he traveled to, where he lived, whom he married, what he wrote, where and when he died.[186]

Narrative always implies a telling of a plot with a determined setting, specific goals, concrete characters, real motivations, etc.—all of which work dynamically to enable the narrative to *go on*, to *keep telling*. It is only through narrative that the apparent independence of plot, setting, goals, characters, and motivations coalesce into a *unitary drama*, a unitary reality, that is, human life itself, each one’s life. Further, as Marías indicates in his example, to speak of human life, it is not a definition that we need, nor a description, but a *narration* of a story, *my* drama—the drama of *my life*. If ‘human life’ is left to the task of simply being defined or described, we would be guilty of committing a violent reduction. When we wish to ask *what* or *who* the human person is, only the *narration* of life’s drama serves as an adequate way of answering those questions and discovering the person as he is in his own life.

The *what* of the human person—or ‘essence’, if you like—is intimately linked to his *who*. We cannot speak of his *what* as an isolated reality from his *who*. The unfortunate separation of these two is clearly articulated by the medieval philosopher Bœthius (and, in fact, many of the Scholastic thinkers who adopted his definition, including Thomas Aquinas) because he thought of the person simply as a substance with a rational nature.[187] *What* I am helps explain *who* I am, just as *who* I am helps explain *what* I am: *soy alguien corporal*, I am a ‘corporeal

[186] Marías, *Metaphysical Anthropology*, 73.

[187] Bœtheius’s famous definition of the person is as follows: *persona est individua substantia rationalis naturæ*, “the person is an individual substance of a rational nature”.

someone'.[188] Without this, we would easily fall into conceiving human life isolated from its fundamental reality: *my* life, *your* life, *his* life, *her* life, and so on, preventing us from narrating anything at all, since narration is executed only discriminately through disjunction: it is about *this* life and not *that* because *this life* is irreducible to *that life* and vice versa.[189]

Now, when we speak of human life, of my life, we speak of it under the function of Ortega's formula: 'I am I and my circumstance'.[190] Therefore, to speak of human life, to narrate human life, my life, my circumstances must never be excluded. To narrate is to narrate me *and* my circumstances. However, here we stumble across an interruption, one that we dealt with previously: it is that human life—because it is systematic, dynamic, and dramatic—is uncertain. (But as we shall see, Marias's idea of narrative is an optimistic approach to facing the constitutive uncertainty of human life). Marías distinguishes 'incertitude' from 'ignorance', the latter being a 'not knowing' and the former being a 'not knowing what to hold by'.[191] To a certain extent, life is characterized by the 'presence' and 'presentness' of these incertitudes, and to overcome them we must be aware of our *situation*[192] and 'give an account' of it.[193]

I must give an account of my situation, *narrate* it, if I am to navigate through any sort of incertitude in my life. For, when I narrate my situation or the drama of my life, when I relate it, when I give an account of it, I find out what I should hold by, and therefore my circumstance acquires relative stability based on a degree of certitude. But, we must not forget, human life—hence my circumstances, too—will always be unstable and precarious: it is constant anticipation of *what will*

[188] Marías, *Metaphysical Anthropology*, 33. See also Raley Harold (1997) *A Watch Over Mortality: The Philosophical Story of Julián Marías*, State University of New York Press. Raley paraphrases Marías's term with much more poeticism, "Someone who is also some-body".

[189] For a more detailed treatment of 'disjunction', Marías explains it in his *Metaphysical Anthropology*, generally in 'Interpretation, Theory, Reason' and 'Empirical Structure', but especially in 'The Sexuate Condition'.

[190] José Ortega y Gasset proposed an understanding of human life in the formulation: 'I am I and my circumstance'. For an elaboration of this Ortegan-Marían metaphysical doctrine, Marías discusses this in his book *José Ortega y Gasset: Circumstance and Vocation* (1970).

[191] Marías, *Reason and Life*, 88.

[192] "The term *situation*, on the other hand, alludes to a much more circumscribed reality; it refers only to those elements of the circumstance the variation of which defined each phase of history and which *situate* us at a certain historical level": Marías, *Reason and Life*, 29.

[193] Marías, *Reason and Life*, 90.

be or *where I will find myself next*.[194] This is also why each human life, each *I*, "tells a story or narrates *for something*, and this sends us to the future".[195] In any case, our concern is that narration sheds light on the situation in which I find myself, and so with more clarity, I more fully find myself *in* it. Thus, the *dramatismo* of my life is, as it were, 'magnified'.

That is why "all thinking, and for profounder reasons all speaking, always occurs with reference to a situation", that is, a circumstance involving certain things.[196] This means that all sorts of narration refer to a context, *living*, "which is the total situation within which [all forms of narration] are given and within which they have meaning".[197] In other words, the things around me receive a personal meaning because they are prerequisites for *living*. That is not to say, however, that things mean only what they mean to the extent that they mean something *to me*. Such an erroneous conception removes from us all responsibility toward the 'outside world'—to anything outside that fundamental reality that is my life—or to whatever is not *me*.

Rather, things are given another layer of meaning—a personal dimension that is relevant to my life—when they *concern me*. "I can find meaning for something only by living," Marías says, "that is, by making it really function within the ambit or area of my life".[198] When I take something to inform my living, then it assumes a deeply personal meaning to my life. Another way of articulating this, coming from what we have said about drama, is to admit that things take on a dramatic quality insofar as they relate to and refer to me. That is why all things surrounding us, surrounding me, can be narrated—*has* to be narrated. Thus, "every vital act… is an interpretation" of the things that I act upon and which move me to act.[199] Some splendid words from Marías sum up what we have

[194] "Human life is not everlasting, but has begun and will end—most important of all, will end, whatever its *ulterior* fate. Furthermore, its possession is not simultaneous, but specifically successive—it is possessed bit by bit—and it is not perfect, but highly imperfect and precarious: unstable in the present *instant*, pale and impoverished in memory of the past, uncertain and vague in anticipation of the future": Marías, *Metaphysical Anthropology*, 210.
[195] Marías, *Razon de la filosofía*, 189, this passage was translated by Paul Dumol (July 2023).
[196] Marías, *Philosophy as Dramatic Theory*, 45.
[197] Marías, *Philosophy as Dramatic Theory*, 45.
[198] Marías, 'The Idea of Metaphysics', 363.
[199] Marías, *Reason and Life*, 186.

expressed and make it possible to grasp them in greater depth: "Life's only mode of *being* is, self-evidently, *living*; and the only mode of speaking about it, in its concrete reality, is *recounting* it",[200] and "this explains why to live is necessarily to give an account (*dar razón*) of what one does in each moment; i.e., to do, in that moment, something specific, in view of the totality of my life".[201]

4. Conclusion: A Preliminary Step to Philosophy

What, then, are we to do with these two concepts? How are they employed in any serious undertaking of philosophy? For one, we can earnestly admit that literature, when it is faithful to the demands of the drama and narrative proper to human life, serves as an entry point into philosophical inquiry. As we noted at the beginning, literature was the first area that took seriously the seemingly mundane fact that the human person is its chief subject matter, the object of its investigation. It is the concepts we have surveyed—*dramatismo* and narrative—that operationalize that characteristic of literature that penetrates and illuminates the reality of the human person, which amounts to literature assuming its place as a precursor to philosophical anthropology. Drama and narrative are not simply literary principles and conceptual tools; they are realities very much present in human life, in each person. Far from being merely abstract notions or theoretical interpretations of human life, they are real—but only insofar as they are manifested in life that is concretely lived, that is, *my* life, the life of each one.

These two concepts, which are indispensable in the area of literature, are key notions in the discovery of the human person. That is why they are essential conceptual motors to reaching philosophy, especially to the analysis and discovery of human life—philosophical anthropology. We have said that human life is constitutively dramatic, and the only way to give an account of life and its drama is to narrate it. Literature, especially novels, are most reflective of this reality. Novels make an account of the drama of a singular life, to such a great extent that Marías said boldly that "man must be the novelist of his own life".[202]

[200] Marías, *Reason and Life*, 194.

[201] Marías, *Reason and Life*, 188.

[202] Marías, Julián (1967) *History of Philosophy*, Stanley Appelbaum & Clarence Strowbridge, trans, Dover Publications, 457.

Being the novelist of our own lives, we must have, among our conceptual tools, an idea of life, because having an idea of life opens up the possibility of narrating life. In some way, then, literature "offers the possibility of prefiguration, the condensing of the experience of life".[203] The reason for this is that literature "is a preliminary stage of the metaphysical investigation of human life, a provisional stage of philosophical thought".[204]

It is recourse to these two concepts that give access to this 'possibility of prefiguration', which is no less than a preliminary step to the philosophical discovery of the person. This is in no way to say that literature takes on the role of philosophy; rather, it is that literature provides the conceptual tools that make the discovery of the human person in human life more philosophically transparent. Marías employs other conceptual instruments that originate from a literary character that could help in justifying the claim we have been making, such as the notion of *ilusión* (loosely, 'hope', 'excitement', 'expectation'). Nevertheless, the notions of *dramatismo* and narrative appear to be the crucial theoretical scaffoldings drawn from literature that build up most effectively toward philosophy. To say that literature, through *dramatismo* and narrative, is a pre-philosophy becomes meaningful if 'pre-philosophy' is understood not as literature's priority in a temporal succession, but as literature's essential—though inchoate—role in discovering the human person and human life.

[203] Cole, Ralph Dean (1974) 'Julián Marías as a Literary Critic', Doctoral Dissertation, The University of Oklahoma.
[204] Marías, Julián (1960) *Obras*, vol 5, Revista de Occidente, 491, my translation.

References

Cole, Ralph Dean (1974) 'Julián Marías as a Literary Critic'. Doctoral Dissertation. The University of Oklahoma.

Marías, Julián (1954) *Ensayos de teoría*. Editorial Barna.

Marías, Julián (1967) *History of Philosophy*, trans. Stanley Appelbaum & Clarence Strowbridge. Dover Publications.

Marías, Julián (1971) *Metaphysical Anthropology: The Empirical Structure of Human Life*, trans. Frances López-Morillas. The Pennsylvania State University Press.

Marías, Julián (1960) *Obras*, Vol. 5. Revista de Occidente.

Marías, Julián (1996). *Persona*. Alianza Editorial.

Marías, Julián (1971) *Philosophy as Dramatic Theory*, trans. James Parsons. The Pennsylvania State University Press.

Marías, Julián (1993) *Razón de la filosofía*. Alianza Editorial.

Marías, Julián (1956) *Reason and Life: The Introduction to Philosophy*, trans. Kenneth Reid & Edward Sarmiento. Hollis & Carter.

Marías, Julián (1967) 'The Idea of Metaphysics', in Aloysius Robert Caponigri, ed. & trans., *Spanish Philosophy: An Anthology*. University of Notre Dame Press.

Mora, José Ferrater (2003) *Three Spanish Philosophers: Unamuno, Ortega, Ferrater Mora*. State University of New York Press.

Raley, Harold (1997) *A Watch Over Mortality: The Philosophical Story of Julián Marías*. State University of New York Press.

Raley, Harold (1980) *Responsible Vision: The Philosophy of Julián Marías*. The American Hispanist, Inc.

AI and the Value of Explanations

Benjamin Robinson[205]

Australian National University

Abstract

AI systems often struggle to explain their outputs. Some have argued that this lack of explainability justifies banning their use in certain contexts. However, this paper argues that in many cases where AI systems are being deployed, the types of explanations we want from these tools can be provided by current methods in explainable AI. I make this case by first distinguishing between the instrumental and non-instrumental reasons why explanations are important. I then apply this analysis to the types of explanations we're able to obtain from AI systems using current explainability methods. This aims to demonstrate that the general, correlative - though not causal - reasons that explainable AI techniques provide are often sufficient for our interactions with large institutions, like governments, hospitals and banks, where explanations are instrumentally important in helping individuals understand, challenge and improve decisions. There are some cases, however, where explanations are non-instrumentally important in that they evidence respect for people as such (end of life care), or where specific causal reasons matter (criminal sentencing), which AI cannot provide. I discuss objections throughout, and finish with the caveat that while explainable AI methods are available in theory, the time and cost it takes to properly implement these techniques means that regulation or other forms of incentives are likely needed to ensure they are actually used in practice.

[205] Ben Robinson is a student in Philosophy at the Australian National University. His work primarily revolves around AI ethics and value.

1. Introduction

AI systems are being increasingly used throughout society; from accessing finance (Booth 2019), to criminal sentencing (Kleinberg et al. 2017), to diagnosing cancer (Yala et al. 2021), to executing drone strikes (Lee 2021). Trained on huge datasets and computing power, these systems are predictively powerful, but we often don't know how or why decisions were made. Unlike traditional algorithms, where rules are pre-specified by human engineers, modern AI systems using machine learning algorithms essentially create their own rules to improve their performance (Mittelstadt et al. 2016), meaning ex ante predictions or ex post assessments of the system's operations are difficult to obtain (Zerilli et. al. 2019).[206]

There are a growing number of examples evidencing the harms of opaque automated decision making,[207] as well as regulatory proposals to enshrine some right to explanation in law.[208] Researchers like Vrendenburgh (2021) have given an account for an individual right to explanation focussing on the rights of decision subjects, while Lazar (2022) gives an account of the duties owed to the political community at large. Some argue that the demand for explainability is overblown; human decision making is just as, if not more, inscrutable than AI because of bias and motivated reasoning (Zerilli et al. 2019). Others argue that we should avoid using AI all together until AI systems can give "full and satisfactory explanations" for the decisions they make (House of Lords, 2019).

While there is something intuitively plausible about the explainability objection, I argue that strong forms of this objection misunderstand both the types of explanations needed in different contexts where AI systems are being deployed, as well as the forms of explanations possible from modern AI systems. In what follows, I aim to demonstrate that methods in explainable AI that give general correlative, though not causal reasons for why a decision was made, are often sufficient for our

[206] Throughout the essay, I refer interchangeably between Artificial Intelligence (AI), AI systems, and automated decision making; and between explainability, inscrutability and opacity. Roughly, AI is any type of computational system that shows intelligent behaviour conducive to reaching goals (Muller 2021). Explainability is the capacity for AI systems to generate intelligible reasons for their outputs (McDermid et al. 2021). Section three of this essay goes into greater detail about both sets of definitions.

[207] In Australia, it was reported that more than 2000 people died after receiving an incorrect government debt notice without adequate explanation or right of reply (Medhora 2019).

[208] For example, Article 15 of the EU's GDPR aims to enshrine some form of explanation in law.

interactions with large institutions like governments, hospitals and banks, where explanations primarily allow for understanding, challenging and improving decisions. However, there are cases, like in criminal sentencing or end-of-life care, where specific causal reasons matter, or where explanations evidence respect for people as such, and AI is unable to provide such explanations. So, overall, the demand for explainability can often, though not always, be met.

To make this case, I first distinguish between different instrumental and non-instrumental reasons why explanations are valuable (section two). I then analyse why AI systems are said to be opaque, and methods to limit this opacity (section three). Section four brings these strands together, where I argue that the instrumental value of explanations - challenging and improving decisions - are generally what's required for decision subjects interacting with large institutions, and these can be provided by techniques in explainable AI. Section five provides a contrary view, arguing that sometimes explanations tracking non-instrumental value - instantiating respect and acting for the right reasons - will be needed, and these cannot be given by AI. I discuss objections as I go, and I finish by showing how the analysis approach outlined in this essay can be used to make sense of individual cases where there is a demand for AI explainability.

2. Why do explanations matter?

I begin by outlining the instrumental and non-instrumental value of explanations, before applying this analysis to AI.

To explain something is to communicate information that enables an audience to reach a justified understanding of it (Wilkenfeld 2014).[209] The mere feeling of understanding isn't enough; it needs to be justified, which could involve explicating reasons, beliefs, intentions, external causes (Malle 2004) or deliberative and institutional procedures that influence a decision (Doshi-Velez et al. 2017). What evidence is drawn upon will be relative to an audience's goals; explanations enabling a data scientist to improve an AI system will be different from explanations

[209] The analysis in this section of what it means to explain something, and the different types instrumental and non-instrumental value of explanations, draws from the approach taken in Lazar (2021) and Lazar (2022).

enabling prediction recipients to understand why their loan was declined. The former will require more technical details enabling a justified understanding of how to improve the internal workings of the system; the latter requires information about how loan decisions are generally made, including factors like savings, debt and credit history. So, explanations serve different purposes in different contexts.

There are at least two broad reasons why explanations are instrumentally valuable. First, they allow decision subjects to challenge decisions and engage in "informed self-advocacy" (Crawford and Schultz 2014; Vredenburgh 2021). If we're told by a government decision making system that we owe the tax or welfare department money, we will want to know how this decision was made so we can challenge it if we think it's wrong. Challenging also applies to other stakeholder groups, like regulators, who require banks to explain how they determine credit scores to ensure compliance with discrimination law. This is linked to accountability, which generally requires that we are able to track who has made certain decisions. It can also identify underlying reasons for decisions, which can enable us to check for things like fairness and discrimination (Barocas and Selbst 2018). Challenging is a basic feature of our interactions with big institutions, be it governments, universities or banks. Even something as innocuous as a parking fine requires a basic explanation so that we can understand whether the rationale was fair, who was responsible, and whether we have grounds to contest.

A second reason why explanations are instrumentally valuable is that they allow for improvements to systems when they go wrong (Lombrozo 2011). If we know that a machine made an incorrect decision, for instance diagnosing a mole as cancerous when it was not, but we don't know why it made that decision, we can't intervene to improve it. This is particularly relevant for data scientists who develop and refine AI systems, but also for the business owners of these systems and regulators. Explanations of this type also allow individuals to understand decisions to improve their chances for next time; if a bank denies a loan because of factors including too much debt, individuals can aim to alter their behaviour before reapplying. So, instrumentally, explanations enable decisions to be challenged, they allow systems to be improved, and relatedly, they allow individuals to interact more effectively with those systems.

There are at least two reasons why explanations are non-instrumentally valuable. First, they are an important part of answerability to others as moral equals; denying someone an explanation for a decision you've made that affects them may constitute a denial of their equal moral standing as a human and mutual membership of a moral community (Lazar 2021). If this is right, even a seemingly banal example of a worker denying a colleague an explanation for why they borrowed their mug or were late to a meeting could be a rejection of their answerability to another as a moral equal. This type of explanation makes most sense in interpersonal settings as it seems to require certain mental states from the actor, namely recognising another as an agent of equal moral standing, and acting from a motivation of equality and respect for them as such. It's not clear that organisations could fulfil this. Organisations can give explanations to decision recipients to fulfil other functions, like understanding and challenging decisions, but recognising another as an agent of equal moral standing, and acting from a motivation of equality and respect for them as such, does not seem possible for groups lacking mental states.[210]

A second reason why explanations are non-instrumentally valuable is that they demonstrate that actions were based on the right kinds of reasons. We often want not just a decision, but reasons behind a decision, where the process of deliberation is itself important (Christopher 1998, from Lazar 2021). Acting on the wrong kinds of reasons can itself be wrong; for instance, if our deliberations were based on sexist reasoning, or on information that should have been private – explanations allow this information to come to light. These sorts of reasons could make a decision unjustified, or less praiseworthy, even if it happened to be the right decision. Consider J.S. Mill's case of saving a drowning man in the hope of being paid a reward (Mill 1863). Individuals are also accountable for their intentions and beliefs; for example, the differences between first and second-degree murder, the first being

[210] This is not to deny that there may be individual people *within* institutions capable of giving explanations of this type. But many of the interactions we have with institutions, whether it be governments or banks, are dealings with them as an entity rather than individuals within this entity. This issue is explored further in the final section of the essay.

premeditated and the second unintentional and thus not as bad as the first, or the role of mental states in establishing recklessness and negligence (Lazar 2021). So, answerability to others, and demonstrating that one acted for the right kinds of reasons, are two ways in which explanations can realise non-instrumental value.

3. Why are AI systems opaque?

Before analysing whether or not AI systems can realise the types of explanation outlined above, it's necessary to first understand why AI systems are said to be opaque, and consider methods to reduce this opacity.

There are at least four forms of opacity in AI systems; institutional, commercial, educational and technical. Institutionally, algorithms are often designed in organisations with input from many engineers and developers over time, meaning "a holistic understanding of the development process and its embedded values, biases and interdependencies" is not possible (Mittelstadt et al. 2016: 7). Commercially, the data and algorithms used in algorithmic systems are often kept secret, backed by trade secrecy protections (Burrell 2016). Educationally, relevant parties might not have the required expertise to understand decision outputs, even if (mathematical or technical) explanations can be given (ibid). However, the concern about AI systems' inability to provide explanations usually relates to the *technical* inability for its outputs to be explained in a way that even an expert would be able to understand. This is what Vrendenburgh (2022) calls "in principle explainability", an issue that is said to afflict only modern machine learning (ML) algorithms.

What makes algorithms technically opaque? Traditional algorithms did not have an explainability problem like those faced by current machine learning systems. This is because in traditional algorithms, rules and weights were pre-specified by the human engineer (Mittelstadt et al. 2016); systems could not do anything that was not already factored into the developers' design for how it should operate given certain inputs.[211] Machine learning, on the other hand, uses vast amounts of data combined with algorithms that can create their own rules to improve their performance. When

[211] Though as Zerilli et al (2019) note: "traditional algorithms, like expert systems, could be inscrutable after the fact: even simple rules can generate complex and inscrutable emergent properties. But these effects were not baked in."

trained on a certain decision task, like whether to approve a loan, these systems "essentially derive [their] own method of decision making" where it is "simply not known in advance what rules will be used to handle unforeseen information" (Zerilli et. al. 2019, 6). As a result, ex ante predictions and ex post assessments of the system's operations are not possible (ibid). But this ability to find patterns in huge amounts of data is also part of their promise; along with being used in decision contexts where humans used to be, for example diagnosing a cancerous mole, they are also being used in new contexts where human decision makers do not have robust, predictively powerful causal generalisations (Barocas and Selbst 2018); like predicting cancer many years in advance.

A fundamental issue with these models is their detection of correlations rather than causation. Machine learning algorithms often find surprising correlations in huge datasets, but the complexity of these algorithms means it is difficult to pick out a smaller set of explanatorily relevant variables and simple relationships between those variables which could explain their decisions (Vrendenburgh 2022). Of course, some of the detected correlations might have causal underpinnings, but the problem is that it's not possible to tell, at least not without detailed further investigation. As Vrendenburgh notes, complexity and an inability to pinpoint causation is not always a barrier to understanding; the natural world is often incredibly complex, but scientists create simplified models to understand it. Yet techniques used to create simplified models in the natural world, like idealisation and abstraction, are not available for these machine learning models because of their complexity.[212] So, while traditional algorithms did not have an in-principle explainability problem, machine learning algorithms do, because of their complexity and their detection of correlation rather than causation. However, there are techniques to limit this in-principle opacity, and so the extent to which opacity matters depends on the kinds of explanations needed in different contexts.

[212] Barocas and Selbst (2018) give four reasons why machine-learning models are complex: linearity, monotonicity, continuity, and dimensionality. Basically, this means that ML models don't act in a linear and comprehensible way, allowing them to identify unintuitive and unexpected correlations, but also causing problems for understanding how they have come to their conclusions.

The field of explainable AI has grown significantly in recent years. Techniques have been developed to improve the explainability of decision-making systems, including by making simpler approximations of a model, or by creating counterfactual explanations that show how a model's prediction can be changed by changing one input. The demand for explainability has also led to certain techniques being prioritised in the model development phase, for instance ensuring that training data are correctly labelled and categorised, meaning that the idea of "black box" AI is less well- founded now than it was in the past. McDermid et al. (2021) outline three types of explainability methods. First, for relatively simple models using linear regressions or decision trees (where there is a direct linear relationship between features of the dataset and outputs), understanding how the model works is straightforward; the weights in the linear regression model can be isolated and extrapolated to give insight into the larger model. Second, for complex ML models, there are feature importance methods, which involve changing an input feature and observing the difference with the original output. For example, if a model's prediction (e.g. for credit risk) does not change much by tweaking the value of a variable (e.g. age), that variable for that particular data point may not be an important predictor. Third, there are example-based methods, where particular input instances are used to explain complex ML models, for example using counterfactual explanations, adversarial examples or influential instances. In the credit risk example, a counter-factual explanation may ask, "if I had more savings/less debt, would I have been approved?" It's not possible to provide exact causal reasons, or the exact weightings of different factors, but general features of a model providing insight on how it made a decision can be given.

To be sure, there may still be many problems with these techniques to increase explainability. For one, these may be quite technical solutions, appropriate for data scientists, but which might need to be translated into terms that other stakeholders, like regulators, business owners, end users or the public can understand. Explainability techniques can also be expensive and time-consuming; businesses developing AI models may not be incentivised to invest in these methods without government intervention. There is also the problem that more intrinsically explainable models (the "simple ML models" that McDermid et al. outline) are generally less powerful and accurate, so there may be a trade-off between explainability and accuracy. But even for the techniques to improve explainability for more complex models (feature and example-based methods), all that is being provided is a correlative explanation, not a causal explanation. The question then becomes whether causal explanations are indeed needed for work contexts where AI is being deployed.

In what follows, I argue that merely correlative explanations provided by explainable AI are sufficient to achieve the instrumental value of explanations in many cases. However, there are work contexts where the non-instrumental value of explanations are important, and therefore AI should not be used given its inability to provide causal reasons, or reasons that instantiate respect for people as such.

4. AI and the instrumental value of explanation

I now move to the core of my argument, where I make two main claims. First, the instrumental value of explanations - challenging and improving decisions - are generally what's required for decision subjects interacting with large institutions, and these can be provided by current techniques in explainable AI. Second, sometimes explanations tracking non-instrumental value - instantiating respect and acting for the right reasons - will be needed, and these cannot be given by AI. So far, I've mentioned various stakeholders to whom explanations may be owed – decision subjects, data scientists, business owners, regulators, the general public etc. From here on, my focus is primarily on decision subjects; while explanations are important to various stakeholders, decision subjects have the most at stake. They are the ones most directly affected by legal, medical, financial and governmental decisions made about them, and so they have the strongest claim to explanations.

My first claim is that the main reason explanations are important to decision subjects is that they allow decisions to be understood and challenged. Most cases of automated decision making, both now and in the near future, involve large institutions making decisions that affect individuals. While we might be worried about AI being embedded within Spotify or Google to recommend music or websites, the demand for explanation is greatest for decisions that significantly affect our life chances. When we're rejected from a job application, denied a loan, recommended a certain health treatment, refused bail etc. we want to know how and why these decisions were made so that we can contest them if need be and navigate these institutions to achieve our aims in the future. There may be non-instrumental value in understanding why the decision was made, which we may care about even if it doesn't translate into action, but primarily explanations serve an instrumental good of navigating institutions; what Vredenburgh (2021) calls "informed self-advocacy". This includes both "forward looking exercises of agency" including understanding rules and procedures of an organisation to achieve one's goals, whether it be getting a loan or accessing welfare or a certain health treatment, as well as "backwards-looking exercises of accountability" to remedy mistakes or unfairness. Explanations have always served this important function of illuminating our social world and interaction with large institutions (bureaucracies are notoriously opaque) but AI ups the stakes and the challenges.

However, this type of explanation that allows for a general understanding of decisions in order to contest, can be provided by AI systems. As outlined above, methods in explainable AI essentially equate explanations with post-hoc interpretability. That is, they allow decisions to be understood after the fact by identifying general relevant factors that influenced the outcome. Post-hoc interpretability allows for enough information to be gathered to be able to contest decisions and engage in informed self-advocacy. Consider the case of a bank using AI to assess loan applications. Here, general information about how loan decisions are made, for instance credit history, amount of debt, the size of the loan relative to amount of savings etc, would be enough for individuals to assess whether they have been treated fairly or if they have grounds to contest. If, for instance, it was shown that age or postcode was factored into the decision making, then this would be grounds to challenge as it may be discriminatory. Individuals might think they have an adequate credit history, or aim to increase their savings and decrease debts before applying again for a loan. What's important here is that all that's required to achieve explanations allowing for contesting or improving decisions are general factors and counterfactual explanations. Causal underpinnings of the decision are not needed for many of our interactions with large institutions.

Now, one might object that institutions owe us specific causal reasons for decisions that seriously impact us. Perhaps causal reasons are necessary to establish why something is unfair, for example if a decision has factored in demographic details like gender, race or age, or why we think the wrong decision was made, like if we think our credit history or savings are adequate. While this seems reasonable in one sense, it would be placing too high a burden on institutions. Even before AI, the right to explanation does not exist in the abstract; it imposes costs on institutions who need to put structures in place so that such explanations can be provided (for example, ensuring information on websites is up to date, that there are adequately trained customer service representatives etc.). Consider acceptance into a university: it is enough to provide applicants with general factors that influenced the decision such as past grades and departmental fit, without detailing specific reasons for individual candidates. The same applies for many other decisions, whether it be applying for a loan or accessing government welfare. And indeed, the counter-factual approach of identifying general factors will allow for things like discrimination to be uncovered; if an institution says that they have factored in age, race or gender, the decision subject can make a judgement of whether this is appropriate. For university admissions, this would be acceptable due to affirmative action considerations, but for loan approvals, likely not due to discrimination law. Maybe in a perfect world we would get specific causal reasons for every decision made about us, but this seems generally unnecessary for the main value explanations provide to decision subjects in navigating institutional rules and procedures.

Another objection asks whether the methods in explainable AI can really provide the types of general-purpose explanations that are needed to understand organisational systems and rules. Instead of identifying general factors that influenced a decision through a counter-factual approach, what is needed is a more holistic overview of the purpose of a decision, how it relates to other decisions and policies within an organisation, and perhaps also mechanisms to contest (e.g. right of appeal services).[213] For people who've had a bank loan application denied, what's needed is not just factors that influenced the decision (savings, credit history etc) but also perhaps some general points about the rationale behind risk calculation for a bank, other services available, where to contest and so on. Current methods in explainable AI are designed to illuminate the technical aspects of a decision tool, rather than to provide such general purpose explanations in natural language.

This seems right, but it just shows that the account so far is incomplete. Explainable AI methods will just be one factor in a bundle of initiatives needed to help individuals understand and navigate institutions. Other factors include general purpose explanations about the purpose of a decision or tool, its contexts, and information about discrimination and fairness, due process etc.[214] Moreover, with advances in natural language processing and text generation, it is entirely possible that these more general types of explanation and information surrounding decisions could be provided by AI. In any case, these are all familiar functions that institutions already have, and the use of automated decision making does not render these factors obsolete, in fact it likely just highlights their importance. So, explainable AI techniques may need to be used in conjunction with more general-purpose explanations so that institutional decision making is understandable and contestable.

5. AI and the non-instrumental value of explanation

I now argue that there are some contexts where the non-instrumental value of explanations appears important, and this won't be able to be achieved by explainable AI. My analysis so far has focussed on individuals interacting with large institutions: using AI to determine eligibility to a good or service, like welfare payments, a job, entry into university, or recommending a course of treatment like in

[213] Two papers that make a similar point are Zerili et al (2019) who draw on Dennett's theory of 'intentional stance' to say that explanations need to be targeted at the level of practical reason, rather than uncovering the architectural innards of a tool. And Vredenburgh (2021), who calls these 'normative explanations' where the purpose is to communicate the normative reasons for why a decision was made, including possibly reference to organisational policies, without reference to casual reasons.

[214] While this does seem to rely on a decision subject having generally high levels of education and awareness, for example about what counts as discriminatory in loan or hiring decisions, this is a general problem for institutions and not specific to AI. It could be countered by ensuring accessible links to relevant legislation, or simple explainers for what people's rights are. As Vredenburgh (2021) highlights, most rights (whether it to be explanation, privacy or whatever) impose costs to certain parties to ensure those rights are upheld. Organisations have always had structures in place to ensure decisions can be understood and challenged by individuals, and these functions arguably just need to be adjusted and strengthened in an age of AI.

healthcare. This is the site of much current automation. But there are interpersonal examples, both now and in the future, that fall outside these cases, where the value of explanation is non- instrumental because it evidences decisions were based on the right reasons, or it instantiates respect for people as such. In these cases, the use of AI appears inappropriate if we want to retain the important non-instrumental value that explanations provide.

Consider explanations evidencing that decisions were made for the right kinds of reasons. We often want not just decisions, but reasons behind a decision; acting on the wrong reasons can itself be wrong. For instance, if our deliberations were based on sexist or racist reasoning, or on information that is private, explanations allow this to come to light. We've seen that some of the underlying reasons for AI decisions can be uncovered, for example through counterfactual approaches to explanation that outline the general parameters for how decisions were made. But the fact that this relies on correlation means it still falls short of being able to give causal reasons for decisions. We may know general factors that influenced an outcome (like a loan application), but we don't know specific reasons. This may be acceptable for interactions with large organisations, where the main value of explanation lies in us understanding in general terms why decisions were made so that we can navigate systems and challenge decisions, but it is more troubling in cases where these underlying reasons matter in their own right.

Consider criminal sentencing – here, it's not just a matter of coming to the right decision, but about judges weighing up and giving specific reasons for decisions, balancing factors including the defendant's intention, knowledge and negligence with impacts on other parties. Even if we could develop an incredibly sophisticated legal program that synthesised all aspects of relevant case law, and was able to come up with judgements that were verified by, say, blind reviews by the most experienced judges in a jurisdiction, not knowing the specific causal reasons for judgements would be a problem. One concern is how a defendant could appeal a decision if they don't know the specific legal argumentation and reasons given. But beyond this, there seems to be non- instrumental value in drawing on the right kinds of reasons. We want judges to be able to prioritise certain kinds of reasons, like intention and negligence, while at the same time disregarding other types of reasons, like demographic details or their intuitions about a person. There is non-instrumental value in knowing and being able to evaluate the reasons that a judge gives, independently from the instrumental values this also achieves in being able to challenge decisions. Similarly, in health care, it may be hard to justify a hospital using AI for end of life decisions, even if it could be guaranteed that such decisions were optimal (e.g. best use of resources), if they can't give specific reasons for why a patient's life ended. There is non-instrumental value in knowing the specific reasons why one sick patient was given an organ donation over another, for instance. In such cases involving interpersonal accountability, corelative explanations are not enough.

This leads to a second reason why explanations are non-instrumentally valuable: that they evidence respect for people as such and are an important part of answerability to others as moral equals. Denying an explanation for a decision you've made that affects someone else could constitute a denial of their inherent worth as a human and mutual membership of a moral community. Here, following Lazar (2021), the idea is that part of answerability to others as moral equals is being responsive to them, and giving them explanations when decisions are made that significantly affect them. These types of explanation are not just about justifying one's actions, but about explaining oneself from a motivation of equal regard for another.

But is this right? Even if we accept that this kind of answerability is important for many interpersonal relationships, it's less clearly needed in many work contexts dictated by rules, procedures and risk management protocols. In many jobs, it's enough that workers perform their basic functions well. The institution as a whole should be able to justify its practices with the right kinds of explanations. But it would be unfair to expect this from each worker. A nurse is required to explain what medicine they're giving to a patient because it's hospital procedure to explain to a patient what treatment they're being given. A bank teller is required to explain to a customer what they've done with their money because it is company policy. There may be underlying reasons for such policies, such as consent or legal liability, which may ultimately be grounded in something foundational like respect for persons, but the reason why workers provide such explanations is because of procedure. Much of work life is dictated by rules and policies, and so even though it would be nice to retain this non- instrumental value of explanation, it's not strictly necessary.

However, there are examples, like the case of determining criminal liability, where these sorts of explanations are constitutive of the good in question. Part of natural justice arguably involves being seen and tried by another member of one's moral community. There is something of value in being tried by a human judge, even if they are biased or coldly rational, because they are part of one's moral community and share similar capacities to us. Similarly, in certain types of life-or-death medical decision making, such as determining which sick patient to give a blood transfusion or kidney transplant, it matters that patients or their families are given explanations from another being within their moral community who can be seen to relate to their plight. Given the moral weight of these decisions, it seems important not just that an explanation is provided about the rationale for such decisions, but that this explanation comes from a person capable of relating to them as an equal. Of course, there might be cases where we choose to sacrifice this value, say if a fully automated hospital would save considerably more lives than a non-automated hospital. But such a situation would incur moral costs.

Again, it might be objected why being seen by another member of one's moral community is really that important. There are various ways of justifying this idea. One route is to say that certain institutions are constituted by human relationships and practices which do not seem amenable to replacement with machines. Pasquale (2018), for instance, argues that types of automation within the legal system threatens the "deliberative governance" that is a foundation of a "just and accountable social order." Another approach is taken by Danaher (2019) who argues that the aggregate effect of widespread automation will atrophy and degrade the bonds of human interdependence that give our lives shape and meaning. It is outside the scope of this essay to argue why being seen by another human is important in certain contexts, but if we accept even a minimal form of this intuition, it appears to place constraints on the use of AI if we are to account for the non-instrumental value of explanations.

Overall, I have argued that in most cases of automated decision making, the main reason why explanations are important to decision subjects is that they realise the instrumental goods of enabling people to challenge decisions and navigate institutions to achieve their aims. However, there are limits to this: in some cases, it appears that explanations have non-instrumental value in that they evidence that decisions were made for the right causal reasons, or that they instantiate respect for people as such. This analysis approach can be used to make sense of individual cases where there is a demand for AI explainability. First, we can ask: what is the value of explanations in this context? I have outlined two instrumental, and two non-instrumental values in this essay, but there may well be more. Once these values have been identified, we can ask: can AI achieve these values? Taking medicine as an example, explainability methods can help identify general factors about a medical diagnosis or treatment recommendation, which allows us to navigate a hospital system and challenge if necessary. But in cases of end-of-life care, explanations that give specific reasons for decisions or that evidence respect for people as such seems necessary, and therefore AI would be inappropriate in such contexts. This basic framework can be applied to many other cases in the same way, whether it be explanations in the legal system, governments or corporations.

A final caveat: as alluded to in section three, there are commercial and institutional reasons why explainability methods may not be used, even if they are available. Namely, these methods are expensive and time-consuming, so companies or organisations wanting to deploy their AI products as soon as possible may forgo implementing them without some external incentive like regulation. The purpose of this paper is to show how, in theory, the types of explanations often needed can be provided by currently existing technology. Whether these will actually be used in practice is a matter of incentives, politics and regulation.

6. Conclusion

Explainability techniques can give a general indication of the parameters for why an AI decision was made. But it can't give specific causal reasons. I have argued that in many cases, the value of explanation lies in the instrumental value of being able to challenge and contest decisions that are made about us. But in some cases, especially in high stakes decision making such as determining criminal liability or end of life care, specific (causal) reasons really matter. In such cases, the use of AI appears inappropriate given its detection of correlation rather than causation, and the fact that machines are not fellow members of our moral community. By clarifying the instrumental and non-instrumental value of explanations, and applying this to AI as it is being deployed in work contexts, I have argued that the demand for AI explainability can often, though not always be met.

References

Barocas, S., Selbst, A. D. (2018). 'The Intuitive Appeal of Explainable Machines', *Fordam Law Review*, 87 (3): 1085–139.

Booth, R. (2019). 'Benefits system automation could plunge claimants deeper into poverty', *The Guardian*, 14 October. Available at: https://www.theguardian.com/technology/2019/oct/14/fears-rise-in-benefits-system- automation-could-plunge-claimants-deeper-into-poverty (Accessed 1 September 2022).

Burrell, J. (2016). 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms', *Big Data and Society*, 3: 1–12.

Christopher, R. (1998), 'Self-Defense and Defense of Others', *Philosophy & Public Affairs*, 27 (2), 123- 141.

Crawford, K., Schultz, J., (2014). 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms', *Boston College Law Review*, 55, 93-128.

Danaher, J. (2019). 'The rise of the robots and the crisis of moral patiency', *AI and Society*, 34 (1):129- 136.

Doshi-Velez, Finale, et al. (2017). 'Accountability of AI under the Law: The Role of Explanation', arXiv:1711.01134.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). 'Human decisions and machine predictions.' *The Quarterly Journal of Economics*, 133(1):237–293.

Lazar, S. (2021). 'The Value of Explanations', manuscript available from author on request.

Lazar, S. (2022). 'Legitimacy, Authority, and the Political Value of Explanations', keynote address to Oxford Studies in Political Philosophy workshop. Available at: https://philpapers.org/archive/LAZLAA-2.pdf.Lombrozo 2011

Lee, K. (2019). 'The Third Revolution in Warfare', *The Atlantic*, September 11. Available at: https://www.theatlantic.com/technology/archive/2021/09/i-weapons-are-third-revolution-warfare/620013/ (Accessed 1 December 2021).

Malle, B. F. (2004). How the Mind Explains Behaviour: Folks Explanations, Meaning, and Social Interaction. Cambridge, MA: *MIT Press*.

Medhora, S. (2019). 'Over 2000 people died after receiving Centrelink robo-debt notice, figures reveal', *ABC News*, 18 February. Available at: https://www.abc.net.au/triplej/programs/hack/2030-people-have-died-after-receiving-centrelink-robodebt-notice/10821272 (Accessed 20 June 2022)

McDermid J., Yan, J., Porter Z., Ibrahim, H., (2021). 'Artificial intelligence explainability: the technical and ethical dimensions', *Philosophical Transactions of the Royal Society*, 379(2207).

Mill, J. S., (1863). Utilitarianism, London, Parker, son, and Bourn.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., Floridi, L. (2016). 'The ethics of algorithms: mapping the debate.' *Big Data and Society*, 16,1–21

Müller, C., (2021). 'Ethics of Artificial Intelligence and Robotics', *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/>.

Pasquale, F. (2018). 'A Rule of Persons, Not Machines: The Limits of Legal Automation', *Maryland Faculty Scholarship*. 1612.

Vredenburgh, K. (2021). 'The Right to Explanation' *The Journal of Political Philosophy*, 30(2):209-229.

Vredenburgh, K. (2022). 'Freedom at Work: Understanding, Alienation, and the AI-Driven Workplace', *Canadian Journal of Philosophy*, 52(1):78-92.

Wilkenfeld, D. (2014). 'Functional Explaining: A New Approach to the Philosophy of Explanation,' *Synthese* 191:14.

Yala, A., Strand, F., Smith, K., (2021). ‘Toward robust mammography-based models for breast cancer risk’. *Science Translation Medicine*. 13(578).

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). ‘Transparency in algorithmic and human decision-making: Is there a double standard?’, *Philosophy and Technology*, 32(4):661–683.

Two-Dimensional Modal Semantics and the Zombie Intuition

Charlotte Senior[215]

The University of Sydney

Abstract

In this paper I discuss David Chalmers' 'zombie argument' in favour of a dualist theory of qualia. I begin by explaining Chalmers' original argument and David Braddon-Mitchell's use of Two-Dimensional Modal Semantics in "Qualia and Analytic Conditionals" (2003) to distinguish two types of conceivability and put pressure on the zombie intuition as proof of dualism. I then critically evaluate Chalmers' defence of the zombie intuition, focussing firstly on his reliance on microphysical properties and secondly on the questionable alignment between his deductive argument for the 'secondary conceivability' of philosophical zombies and actual experience. Stephen Yablo's account of conceivability, specifically regarding undecidability, is fruitful when explicating the latter issue. I end by concluding that Braddon-Mitchell's two-dimensional account can make sense of the primary conceivability of philosophical zombies, and Chalmer's defence of the zombie intuition at least partially relies on problematic argumentation. However, there are further problems elicited by Chalmers' paper surrounding which structures necessitate qualia, meaning that the problems relating to physicalism his defence presents still bear significant weight.

[215] Charlotte graduated last year from the University of Sydney with an Arts Degree (double major in Philosophy and Anthropology), with honours in Philosophy. Her primary areas of interest are epistemology and metaphysics, of which she took a particular interest in during her final year of study. Charlotte's honours thesis focused on hyperintensionality and its relation to theories of impossible worlds.

1. Introduction

The zombie intuition is the purported conceivability of an exact physical copy of yourself that lacks conscious experience. This physical copy is utterly identical to you in terms of physical composition, meaning that the placement of every atom (or your preferred physical fundamental) is identical, and nothing has been added beyond your exact composition of atoms. The zombie intuition arises from the "epistemic gap" dilemma, whereby there is no full explanation of how conscious experience, or qualia, arises from physical brain states.[216] To be clear, the "epistemic gap" dilemma does not posit that we cannot theoretically give a full explanation of all the physical facts that *produce* qualia, but rather that we cannot comfortably say these physical facts *constitute* qualia. Frank Jackson's black and white room thought experiment illuminates the point well: if qualia is nothing over and above a collection of physical facts, then the physical facts associated with a certain qualitative experience (for example, seeing red), should constitute the experience itself.[217] At the very least, this is the most direct conclusion to draw if physical facts are one and the same with qualia, which in turn seems to be straightforwardly derivable from the proposition that physical matter is all that exists.

It is this proposition which is at the core of *physicalism*, which I shall use interchangeably with *materialism*. Both state the thesis that physical facts, or matter, is all that exists and thereby involved with the production and constitution of qualia. Physicalism, or materialism, is in opposition with dualism, which ultimately posits there is *something more* than physical facts in the production and constitution of qualia.

I intend to focus on the analytic functionalist account of qualia as the core physicalist opponent to dualism. Thus, before I proceed, it is necessary to outline analytic functionalism as a thesis and clarify its place in the following arguments. The view holds that qualia is identical to the physical thing(s) or process(es) that

[216] Chalmers, David (2009) 'The Two-Dimensional Argument Against Materialism', in A. Beckermann, B.P. McLaughlin, and S. Walter, eds. *The Oxford Handbook of Philosophy of Mind*: 313. Oxford University Press.

[217] Jackson, Frank (1986) 'What Mary Didn't Know', *The Journal of Philosophy* **83**: 291-95.

plays the relevant *functional role*; this is what qualia is by definition.[218] The view takes this identity relation at face value and in that sense is the most straightforward version of the physicalist thesis to contrast to dualism. Analytic functionalism places no weight on exactly *what* physical facts constitute qualia, just that qualia is by definition whatever physical phenomena plays the right functional role(s).[219]

Because it is merely the functional role that matters, different qualitative experiences (for example, the experience of seeing red and that of tasting chocolate) could have completely different physical facts constituting them. Hence, it is not necessary for there to be anything consistent underlying all conscious experience.[220] Further, philosophers such as David Lewis (1990) and Lawrence Nemirow (1990) have maintained that qualia is whatever physical phenomena plays the right functional role whilst adding that the *visual representation* of red and all of the *physical facts* that define the subjective experience of seeing red are different 'modes of presentation' of red, whilst both being identical with the subjective experience of red.[221] I mention these details to elucidate the focus. These concerns are supplementary to the foundational claim that qualia is identical with the physical facts that play the relevant functional role. It is in this capacity that analytic functionalism will henceforth represent physicalism and oppose dualism.

Dualists have used the "epistemic gap" as the foundation from which to introduce the conceivability argument against materialism. The argument runs thus: firstly, you can conceive of a philosophical zombie in the sense described earlier. Secondly, if something is conceivable, then it is metaphysically possible. (Henceforth, metaphysical possibility will be defined according to possible worlds. That is, a proposition is metaphysically possible if it is true in at least one possible world, where a possible world is a complete explanation of a way the

[218] Braddon-Mitchell, David (2003) 'Qualia and Analytical Conditionals', *The Journal of Philosophy* **100**: 120.

[219] So, there could be vastly physical facts accounting for different qualitative experiences.

[220] Of course, this may still be the case contingently.

[221] See: Lewis, David (1990) 'What Experience Teaches', in W. Lycan ed. *Mind and Cognition: A Reader*: 499-515.

world could have been. Everything that is metaphysically possible is logically possible, however there are propositions that are logically possible but not metaphysically possible.[222] Regrettably, brevity does not permit a further discussion on the consistency of this term within this paper.) Thirdly, if a physical duplicate of myself that lacks conscious experience is metaphysically possible, then I must have something non-physical accounting for my conscious experience. Therefore, materialism is false.

Chalmers (2009) has formalised this argument. Let Q be the fact that someone has qualia. Let P be the conjunction of all material facts about the world. For instance, P could include the proposition "the door to my house is red". This proposition can also be analysed in terms of more granular physical facts, perhaps about the chemical proposition of the paint, or even the placement of each and every atom (or physical fundamental). What matters is these facts detail the physical state of the world, and making up P, account for all physical phenomena.

The argument runs thus:

(1) P&~Q is conceivable

(2) If P&~Q is conceivable, P&~Q is metaphysically possible.

(3) If P&~Q is metaphysically possible, materialism is false.

(4) Therefore, materialism is false.[223]

The most controversial of these premises is the second. Philosophers have challenged the idea that conceivability entails metaphysical possibility, particularly Yablo (1993) who gives a fruitful definition of conceivability that escapes the binary between conceivability and inconceivability, and Braddon-Mitchell (2004), who targets the zombie intuition and explains its existence using two-dimensional modal semantics. Chalmers (2009) defends the second and third premise by claiming it is valid to infer possibility from

[222] "The car is red but not coloured" is an example of a proposition that is logically possible but not metaphysically possible.

[223] Chalmers, 'The Two-Dimensional Argument Against Materialism', 314.

conceivability in the case of the zombie intuition, because this intuition is a special class of conceivability.

In this paper, I will start by outlining Braddon-Mitchell's account for the zombie intuition in terms of two-dimensional semantics. I will then sketch out Chalmers' response that rests on the notion that Braddon-Mitchell has explained the wrong kind of conceivability, and the right class of conceivability regarding the zombie intuition can be deductively proven. I will then critically analyse the course Chalmers takes to reach his conclusion, specifically his distinction between and use of 'intrinsic' and 'structural' profiles. Finally, I will critique the implication from this argument that we have so called 'secondary conceivability' of philosophical zombies and argue that this conclusion does not align with direct experience.

2. Two-Dimensional Account of the Zombie Intuition

In his paper "Qualia and Analytic Conditionals", David Braddon-Mitchell gives an explanation for the zombie intuition that does not undermine materialism. The basis of the argument is that conceivability and possibility claims are always made on the basis of what we believe is actual. Thus, our assessment of a world considered as counterfactual, in terms of the properties it possesses, is largely dependent on what we believe the actual world is like. If the actual world is one in which qualia is non-physical and dualism is true, then it is an *a posteriori* necessity that qualia is non-physical, meaning this is both necessary and empirically discoverable[224]

Perhaps the notion of *a posteriori* necessity is best explained with an example using a physical phenomenon we encounter everyday: water. We know that water is H^2O and would classify this as a necessary truth, because there is no other chemical composition that could be water. Yet, we did not arrive at the truth that water is H^2O via reasoning (as we do with mathematical truths, for instance), but through experimentation and more broadly, *experience.* The role of experience, rather than pure reason, in uncovering this necessary truth makes it an *a posteriori*

[224] Braddon-Mitchell, 'Qualia and Analytic Conditionals', 120.

necessity that water is H^2O. Similarly, if scientists discovered a full explanation of qualia as physical brain states, the proposition "qualia are physical brain states" would be an *a posteriori* necessity.

The notion that we can empirically uncover a necessary truth is intriguing. Pursuing the water example, an understanding of 'water', or the ability to use the term in coherent conversation, does not seem to be dependent on the knowledge that water is H^2O. Water maintains the same *purpose* and defining attributes independent of this knowledge: it is a clear, potable liquid found in rivers and lakes and so on. Adjacent to this point, we could, *prior* to learning that water is H^2O, conceive (imagine, without apparent inconsistency) of water being another chemical compound.[225] Call this compound XYZ and imagine that when lakes and rivers are filled with XYZ, it is indistinguishable from H^2O. From our perspective, knowing that water *is* H^2O, a possible world with XYZ does not have water but something different. However, if we start from a perspective that lacks knowledge of the chemical composition of water, the XYZ world *does* have water. It seems we can conceive of water being XYZ if we suspend our knowledge of water being H^2O, but so long as we maintain that water is H^2O, the XYZ world is inconceivable.

Let us now develop this line of thinking by introducing the concept of an *indexical*. It is essential to the concept of qualia that it allows an agent to access the intrinsic nature of their experience, and because of this, if an agent believes they have qualia, they cannot be mistaken.[226] It may be useful to compare the statement "I have conscious experience" to the statement "I am here". The definition of "here" is dependent on the subject speaking the utterance; thus I cannot produce an incorrect proposition when I say "I am here", provided that by "here" I mean where I am situated. Similarly, it does not seem possible for a subject to claim they have qualia and make an incorrect proposition. In this sense, qualia is an indexical concept.

[225] In the final section of this paper I will investigate the notion of being able to 'conceive' of something in this sense.

[226] Braddon-Mitchell, 'Qualia and Analytic Conditionals', 123. Comparatively, I can be mistaken in claiming that I have knowledge of a certain subject matter.

Bringing this together, we can now say that if dualism is true in the actual world, and qualia has an essentially non-physical element, then qualia being more than physical facts is an *a posteriori* necessity. Hence, when we conceive of any counterfactual world that lacks non-physical properties, that world does not have qualia. Yet, if the actual world is physicalist, it does not transpire we lack qualia. Because qualia is an indexical, so long as we claim to have qualia, we cannot make an incorrect proposition.[227] From the perspective of physicalist worlds considered as actual, any counterfactual dualist world with non-physical phenomena does not have exclusive access to qualia or even a heightened version of it; they simply have a strange, non-physical additive. So, when dualist worlds are considered as actual, qualia is necessarily non-physical, and when physicalist worlds are considered as actual, qualia is necessarily physical.

Before we proceed, one clarification needs to be made. Concordant with the "epistemic gap" dilemma, philosophers defending dualism do not deny that neuroscience provides *markers* of conscious experience, or qualia. For instance, qualia could be accompanied by brain state ∂ (and a causal relationship between ∂ and qualia could even be produced). However, dualists still maintain brain state ∂ does not constitute the *intrinsic nature* of qualia; it is simply a contributing causal factor to something inherently non-physical.[228]

All of this can be practically elucidated with the help of two-dimensional modal semantics, which effectively distinguishes *a priori* and *a posteriori* necessity and their use in language and reasoning. The following will use tables which visually portray these distinctions. The left column are worlds considered as actual, whilst the top row are worlds considered as counterfactual (from the point of view of a world which is considered as actual).[229] We have four worlds: in the first, qualia bears a causal relationship with brain state ∂, but is defined by extra non-physical phenomena. The second world is the same as the first, except the marker for

[227] The indexicality here only extends to making the claim about *yourself* having qualia – not other people.

[228] Braddon-Mitchell, 'Qualia and Analytic Conditionals',117-119.

[229] To use the water example, the actual world is one in which water is H20, however I can counterfactually consider a world in which there is something that looks like water called XYZ.

qualia is brain state ß. The third and fourth worlds are physical worlds, where qualia are brain states ∂ and ß respectively.

(N) *"qualia are non-physical states"*

	∂ + non-physical phenomena	ß + non-physical phenomena	∂	ß
∂ + non-physical phenomena	T	T	T	T
ß + non-physical phenomena	T	T	T	T
∂	F	F	F	F
ß	F	F	F	F

We can see that (N) is considered true only when the actual world is one in which qualia is inherently non-physical. For these worlds, (N) holds true even for counterfactual worlds without non-physical properties, as in these worlds, there is no qualia. In physicalist worlds, when we consider a counterfactual world with non-physical states, these are not qualia, but some esoteric additional property.

If we were to consider the statement "qualia are functional brain states", the table for this proposition would be the opposite of the one above. From the point of view of worlds with non-physical qualia, qualia are non-physical states in *all* worlds considered as counterfactual; and from the point of view of worlds with functional brain state qualia, qualia are functional brain states in *all* worlds considered as counterfactual. Again, we form our conception of qualia based on the world we believe to be actual. This two-dimensional account explains our concept of qualia, whilst also maintaining its malleability subject to what we believe to be actual.

We can now use this framework to account for the zombie intuition. Observe the table for

(Z) *"zombies are possible"*

	∂ + non-physical phenomena	ß + non-physical phenomena	∂	ß
∂ + non-physical phenomena	T	T	T	T
ß + non-physical phenomena	T	T	T	T
∂	F	F	F	F
ß	F	F	F	F

If we draw a diagonal line from the top left to the bottom right, we have the *A-intension* of this statement. The A-intension accounts for our *a priori* intuition, when we have no prior information on what qualia is (whether it is physical or non-physical) and hence do not know what world is actual. Each of the four horizontal rows are different *C-intensions*, which account for the intuitions we have based on what we think is actual. In this table, the A-intension is contingent, but each C intension is necessary; that is, either necessarily true or necessarily false. A necessary C intension means it is *metaphysically necessary*, given that the actual world is such a way, that the proposition must be true or false. A necessary A intension means that it is an *a priori necessity* that the actual world must be such a way, regardless of a specific detail about the actual world. Tying this back to intuitions, we can imagine that if I am asked if zombies are possible in a ∂ world, and I am certain I am in a ∂ world, then I will confidently respond "no", provided my intuitions are reasonable. However, if I am unsure if I am in a physicalist or dualist world, I will not be able to give such a confident answer. My intuitions will be unclear. If the actual world is a non-physical-qualia world, then zombies are necessarily possible, and worlds ∂ and ß are instances of such worlds. In physical-qualia worlds, zombies are necessarily impossible, and worlds such as *∂ + non-physical phenomena* and *ß + non-physical phenomena* have a foreign additional property that is not qualia.

When we intuit that a philosophical zombie is conceivable, this represents an underlying uncertainty in the truth of physicalism. Even the staunchest physicalist can recognise that if it was discovered tomorrow that non-physical phenomena is intrinsic to qualia, then it would be an *a posteriori* necessity that qualia is non-physical. On the other hand, if we discovered tomorrow that physicalism is true and qualia is ∂ states, then *this* would be an *a posteriori* necessity. Thus, the A-intension represents our concept of qualia with an indeterminate account of what is true actually, whilst the C-intension accounts for our concept of qualia on a pre-set assumption of the way the world is actually. When we conceive of a philosophical zombie it is in virtue of the contingent A-intension. Our ability to conceive of philosophical zombies is in virtue of the credence we give, however small, to dualism being true. The work of determining which world is actual and the (non)existence of non-physical states can be left to the evidence that neuroscience provides, but regarding the conceivability of P&~Q along the A-intension, this is not a sufficient basis from which to draw ontological conclusions.

3. Chalmers' Response

In his paper "The Two-Dimensional Argument Against Materialism", Chalmers defends the zombie intuition, that is, the assertion that we can conceive of an exact physical copy of ourselves that lacks qualia. He does this by claiming that in the instance of the conceivability argument, it is appropriate to infer ontological conclusions from epistemic premises. Chalmers defines conceivability along the A intension as *primary conceivability* (1-conceivability), which defines what is conceivable on an *a priori* basis. Conceivability along the C intension is *secondary conceivability* (2-conceivability).[230] He concedes that "primary conceivability does not entail metaphysical possibility", which Braddon-Mitchell's argument was instrumental in demonstrating.[231] Metaphysical possibility can (very briefly) be understood as ways things could have been in different possible worlds. Metaphysical possibility and metaphysical necessity is encompassed by logical possibility and logical necessity. To return to our water example, although I can conceive of water not being H^2O along the A-intension, it still has a necessarily false C-intension if the actual world is one in which water is H^2O. There is also a distinction between *prima facie* and *ideal* conceivability, where ideal conceivability is free from human cognitive limitations.[232] When Chalmers makes claims regarding conceivability and possibility, it is understood in terms of *ideal* conceivability. Chalmers posits that although we cannot entail metaphysical possibility from primary conceivability, we *can* entail "primary possibility" (1-possibility) from ideal primary conceivability.[233] We can utilise the H20/XYZ example to understand how something can 1-possible but not 2-possible. When I "conceive" of a world in which water is not H20 but XYZ, I conceive of a world in which something with identical qualities to water is XYZ instead of H20. Hence, it is 1-possible for water to be something other than H20.

[230] Chalmers, 'The Two-Dimensional Argument Against Materialism', 317.

[231] Chalmers, 'The Two-Dimensional Argument Against Materialism', 317.

[232] Chalmers, 'The Two-Dimensional Argument Against Materialism', 315. I have concerns about the relevance of ideal conceivability and its relation to possibility, if ideal conceivability is abstracted away from human ability. Arguably, human ability and cognitive constraints are intrinsic to what conceivability is. Unfortunately, these concerns are tangential to this paper.

[233] Chalmers, 'The Two-Dimensional Argument Against Materialism', 318.

Even if ideal primary conceivability entails primary possibility, the zombie intuition is still firmly separated from metaphysical possibility. In order to bridge the gap between metaphysical and primary possibility, the 1-possibility of P&~Q must entail 2-possibility. This would require the A and C intensions of P&~Q to be the same. In the case of qualia, there is a strong case for the A-intension and C-intension being identical, due to the indexicality of the concept.[234] That is, there is no distinction between the intrinsic nature of qualia and it's appearance/function.

Unfortunately for Chalmers, for any concept about physical objects, this is less convincing. We have already identified that the A and C intensions of water are distinct and from this example it appears that P must have distinct A and C intensions. However, rather than giving more examples about physical objects, Chalmers proceeds by focusing on "microphysical properties". Take, for instance, the physical property of acceleration. Chalmers holds that there could be possible worlds where instead of acceleration being instantiated, there is 'pseudo-acceleration', which fulfills the same function as genuine acceleration but is somehow *not* acceleration. He appears to take the comparison between these microphysical properties and their pseudo counterparts as directly comparable to the case of $H^{2}0$ and XYZ.[235] As with physical phenomena, Chalmers holds that the A-intension of this property identifies the *role* this property plays, whilst the C-intension identifies what actually plays that role.[236] It is my stance that this is an inaccurate view to be taken of microphysical properties, and has bizarre ontological implications. However, I will proceed to demonstrate Chalmers' argument before making further criticisms.

For example, take three worlds where 'acceleration' is instantiated by properties A, B and C. If you ride your bike down a hill in property A world, the acceleration, as in property A, is instantiated and makes your bike speed up, and so on with property B and C worlds.

[234] Chalmers, 'The Two-Dimensional Argument Against Materialism', 320. I.e., compare to the $H^{2}0$/XYZ case.

[235] Chalmers, 'The Two-Dimensional Argument Against Materialism', 320.

[236] Chalmers, 'The Two-Dimensional Argument Against Materialism', 321.

K: *"The bike accelerated"*

	Property A	Property B	Property C
Property A	T	F	F
Property B	F	T	F
Property C	F	F	T

The A and C intensions of K are different. K has a necessary A-intension but contingent C intensions. Additionally, this demonstrates that a world can *verify* (play the right functional role) K without *satisfying* K. To expand, if we analysed the proposition "the bike sped up in an exponential fashion", there would be a T in every square. If we take property B world to be actual, a world satisfies K if acceleration occurs as instantiated in property B. A world merely verifies K if it fulfills the same role but is not the same property. Chalmers proceeds by reasoning that if microphysical properties do not have the same A-intensions and C-intensions, the same must be said for P.

Let us bring this back to the zombie intuition and recall our present state. P&~Q is conceivable along the A-intension (1-conceivable) because there are some worlds where qualia are non-physical states and can therefore be absent, even if P remains the same. P&~Q is only 2-conceivable under the assumption that dualism is true in the actual world. Having recognised that the conceivability argument cannot be redeemed by directly proving that the A-intension and C-intension of P coincide, he proceeds from this basis to propose to deductively prove that the distinction between the 1-conceivability and 2-conceivability of philosophical zombies entails the falsity of materialism.

Firstly, in virtue of P having different A and C intensions, a world may *verify* P without *satisfying* P. For any statement about a physical property, a world may have this property, even though what actually plays this role is different. Put practically, according to Chalmers, there could be a world that has the property of mass in the sense that it performs all the same functions as mass in the actual world, except instead of these functions being performed by property M, it is performed by property N. Worlds with "pseudo-mass" therefore have the same "structural profile" as the actual world, but possess a different "intrinsic profile".[237] This can be extrapolated to P more generally: there are worlds that appear the same as ours, in the sense that all microphysical properties are fulfilling the same role, but where these microphysical properties are different to the ones instantiated in the actual world. That is, these worlds have the same structural profile, but different intrinsic profiles to our world. Because P has different A and C intensions, so does P&~Q. Therefore, structural properties *alone* do not necessitate the existence of qualia. Chalmers takes these "intrinsic properties" and asks what work they can be doing to necessitate the existence of qualia beyond mere structural properties. Philosophical zombies must be missing some intrinsic properties that they do not have and that we, in virtue of our qualia, must have.

Chalmers argues that this proves the falsity of materialism, as the distinction between the A and C intension of microphysical properties necessitates a distinction between their "structural" and "intrinsic" profiles, which thereby demonstrates that structural properties alone do not produce qualia. From this we can deduce that materialism is false and hence we can infer the 2-possibility of P&~Q from 1-possibility, or else that Russellian Monism is true. Very briefly, Russellian Monism posits that there are properties that underlie the structural account of physics, which describes fundamentals and properties only in terms of what they *do,* not what they are. These underlying properties are (wholly or in part) constitutive of qualia. The notion that qualia is inherent in the structural account of physics, and hence physical objects potentially possess some of these underlying properties ("proto-qualia"), means that it is often not accepted as genuine physicalism.

[237] Chalmers, 'The Two-Dimensional Argument Against Materialism', 321.

4. Criticism of Chalmers' Argument

4.1 Reliance on Microphysical Properties

I agree with Chalmers that physical phenomena such as water have different A and C intensions. By this I mean that whenever we make a proposition about the intrinsic nature of water (for example, "water is $H^{2}0$"), this proposition has distinct A and C intensions. What I find problematic is his move from distinguishing the A and C intensions of physical *objects* to what he calls "microphysical *properties*" such as charge, mass and acceleration. Ultimately, I do not think it is possible to distinguish the A and C intensions for microphysical properties. The pressing question is whether this clears up the path for P&~Q having identical A and C intensions, thereby falsifying materialism via an alternate route.

Physical objects supervene on physical fundamentals. Whatever we call "water" in the actual world supervenes on the physical makeup of molecules in the $H^{2}0$ compound. The supervenience relation between a microphysical property (take mass as an example) and physical fundamentals works a little differently. Microphysical properties are necessarily *properties* held by physical objects and are defined by the roles they perform *in virtue of* the physical makeup of physical objects. For instance, a brick supervenes on physical fundamentals and has a mass of 3kg in virtue of this supervenience relation.

Let us expand on this. Microphysical properties track *relations*, but they are not things in and of themselves. Mass exists *in* the supervenience relation between physical objects and physical fundamentals. Thus, mass tracks a certain identity relation between physical objects and physical fundamentals (the mass of a whole is defined by the amount of physical fundamentals comprising it). Acceleration seems to be more complex, as we have to consider mass and its relation to gravity, which are other microphysical properties. However, I posit that for each microphysical property, we can plot a path that tracks back to a *relation* of some kind that pertains to physical phenomena, whether that be supervenience relation between physical fundamentals and physical objects, or something else (perhaps between physical fundamentals themselves). If we accept that microphysical properties are defined by relations pertaining to physical phenomena, they cannot be 'performed' by a different property.

Building from this point, even if physical fundamentals were entirely different, there is nothing to necessitate that mass would not exist. As long as there is a property we can assign to objects that differentiates the quantity of matter in a physical body, the actual constitution of this matter is irrelevant. Therefore, it is incoherent to say that mass is "performed by property M", and thereby to distinguish distinct A and C intensions. Mass cannot be performed by another property because mass is itself a property.

Thus, I believe the A and C intensions of microphysical properties are necessarily the same. Chalmers does acknowledge this view in his paper, however he does not expand on it beyond a singular mention. To quote Chalmers:

> There are other views of the semantics and metaphysics of microphysical terms that may reject this argument for the distinctness of the primary and secondary intensions of 'mass'. In particular, the argument will not go through on views according to which it is necessary that mass is the property that plays the mass role. … Still, the view sketched above is a quite reasonable view—more plausible than the alternatives, in my opinion— and it is the view that best grounds resistance to an inference from the 1‑possibility of P&~Q to its 2‑ possibility. (Chalmers 2009: 321)[238]

Hence, Chalmers offers no defence of his view of microphysical properties and why it is "more plausible that the alternatives". The problem remains unaddressed by Chalmers: how can the role of a microphysical property be distinct from the property itself?[239]

[238] Chalmers, 'The Two-Dimensional Argument Against Materialism', 321.

[239] A quick side note: Chalmers' idea that microphysical *properties* have a distinct A and C intension (or intrinsic/structural profile) could display an inherent bias for dualism, or at the very least, a bias against analytic functionalism, as this view separates the property from its functional role. If Chalmers finds it intuitive to make this distinction with microphysical properties, the distinction feels a lot easier with qualia. This is merely an observation and not a critique of the formal argument.

If the A and C intensions of microphysical properties are the same, this could disrupt the original argument. P can be defined as the conjunction of all the microphysical facts in the actual world.[240] If we were to define a physical fundamental (such as atoms) and provide a complete account for their organisation throughout the universe, then microphysical properties would arise out of that. That is not to say *all* microphysical properties would necessarily arise out of any configuration of fundamentals, because what matters is having the right *relations* between physical fundamentals (it may be possible to have a world without mass or charge, but this is a separate issue). Perhaps we can say there are a series of ways the world could be constituted and organised so as to produce certain microphysical properties, and this is dependent on there being the right relations pertaining to physical phenomena. In the set of worlds where we have mass, for example, mass is the same property in every one of them. Therefore, on my view the A and C intensions of physical objects are different, but the A and C intensions of microphysical properties that arise out of physical objects are the same. No matter how physical objects are constituted and organised, this will have no impact on the mass, charge and acceleration in a world, provided these properties exist in this world. Comparatively, this is not the case with physical objects. Even if XYZ plays the water role in Twin Earth, it is not water, provided $H^2 0$ is water in the actual world.

Here is one way to close this argument in favour of materialism: Provided that P is the conjunction of all microphysical facts, and microphysical properties simply arise out of that, we can maintain that the A and C intension of P are distinct, and hence the A and C intension of P&~Q are distinct. So, the deductive argument does not go ahead. However, it must be granted that no explanation has been given for exactly *how* microphysical properties arise out of physical phenomena. Do they arise through the structural relations, intrinsic relations or both? I have already intimated that they arise out of the structural relations *alone*, because if the intrinsic relations are involved then we have the same problem, that is, that the

[240] Chalmers, 'The Two-Dimensional Argument Against Materialism', 314. Whilst I gave a broader definition of P earlier as the conjunction of all physical facts, here I am simply specifying microphysical facts as the fundamental facts about fundamentals which comprise the physical facts.

complete structural account of physics does not account for microphysical properties. But if intrinsic relations have no position in the creation of microphysical properties, what *work* are they doing at all?

Perhaps the apparent redundancy of intrinsic relations produces the same problem with relation to qualia. There could be a world W where P is verified along the A-intension but not satisfied. This world may have XYZ instead of $H^2 0$. It also may have 'pseudo' versions of other physical things (perhaps oxygen, atoms, etc.). This world could have the same microphysical properties, such as mass, charge etc. This world has the same structural relations but different intrinsic relations. Is it necessary that this world has qualia? If it is not necessary, then the structural account of physics alone does not account for qualia, so the 'intrinsic properties' must be doing some work. However, if it is necessary that this world has qualia, then intrinsic relations are again made redundant. Hence, the same problem is recreated through slightly different means, suggesting that the two-dimensional argument against physicalism still bears some force.

4.2 Problematisation of 2-Conceivability

Along with positing a link between 1-possibility and 1-conceivability, Chalmers is also consistent in extending this link to exist between 2-possibility and 2-conceivability.[241]

The conclusion for Chalmers' argument is "materialism is false or Russellian Monism is true".[242] The preceding section criticised the passage taken to this conclusion, however if we discount Russellian Monism for this instance and assume the truth of dualism as the only alternative to materialism, then we must have 2-conceivability, also known as *direct* conceivability of philosophical zombies. From first sight, the idea that we have direct conceivability of philosophical zombies is highly questionable in virtue of the continued existence of this very debate. Whilst Chalmers may claim to have proven that direct conceivability of zombies is possible, the same confidence is not present in our

[241] Chalmers, 'The Two-Dimensional Argument Against Materialism', 318.

[242] Chalmers, 'The Two-Dimensional Argument Against Materialism', 322.

own self-evaluation. Precisely what are we conjuring up when we 'conceive' of a philosophical zombie? I will endeavour to provide some thoughts in response to this question with the aid of Stephen Yablo's work on conceivability. I have chosen to focus on Yablo's work as it is effective in formulating an account of conceivability that escapes the binary between conceivability and inconceivability and thereby produces, in my view, a satisfying account of the subjective experience of 'conceiving' of something that may not be possible.

Stephen Yablo, in his paper "Is Conceivability a Guide to Possibility?" defines the conceivability of proposition p (CON) as being able to *"imagine a world I take to verify p"*. Inconceivability (INC) is defined as being unable to *"imagine any world I don't take to falsify p"*.[243] Note that upon this account, inconceivability is not simply the negation of conceivability. In fact, the negation of conceivability, (*"I cannot imagine a world that I take to verify p"*), and the negation of inconceivability, (*I can imagine worlds that I don't take to falsify p*), are not mutually exclusive.[244]

Importantly, the account of conceivability as (CON) is in terms of worlds and not just singular situations. In my view, this is linear with our regular experience of conceiving. When I conceive of a proposition, say, (E) "there is a pink elephant waiting for me in my lounge room", I am in one sense imagining a singular situation, however the wider world is implied. I may not imagine all of the details, (for example, the exact shade of pink and the placement of the elephant's feet), but this does not mean I imagine the situation to be abstract in these ways. When I conceive of (E), the elephant I conceive of *does* have a specific shade of pink and exact placement of the feet. Further, if it was not possible to specify these details without avoiding incoherence, then (E) may not be conceivable anymore. In this sense, I imagine the situation *as determinate*, but because I do not spell out every single detail in my conceptualisation, I do not imagine it *determinately*.[245] In the same way, this account necessitates that every proposition that is conceivable

[243] Yablo, Stephen (1993) 'Is Conceivability a Guide to Possibility?', *Philosophy and Phenomenological Research* 53: 29.

[244] Yablo, 'Is Conceivability a Guide to Possibility?', 31.

[245] Yablo, 'Is Conceivability a Guide to Possibility?', 28.

entails an entire world, because in theory, it should be possible to detail everything about the world where this proposition is true.

Further, under Yablo's account, there is room for the zombie intuition to be undecidable, that is, neither conceivable nor inconceivable. We can potentially have the conjunction of these two statements, which are the negation of (CON) and (INC):

> (~CON) I cannot imagine any world in which P&~Q is verified.
>
> (~INC) I can imagine worlds in which it is not falsified that P&~Q.
>
> (~CON & ~INC) I cannot imagine any world in which P&~Q is verified, and I can imagine worlds in which it is not falsified that P&~Q.

The conjunction of these propositions could entail there are worlds where P&~Q is in between true and false, or else its truth value is ambiguous. That is:

> (a) I can imagine a world in which P&~Q has a truth value that is in between true and false.
>
> OR
>
> (b) I can imagine a world in which P&~Q is not verified, but it is also not falsified, because the situation is ambiguous.

In order to avoid a discussion on non-classical logic, we shall proceed with interpretation (b) of undecidability.

This interpretation of undecidability can also be elucidated in terms of two-dimensional semantics and Chalmers' primary and secondary conceivability. Suppose that a proposition is 1-conceivable if there is at least one T in its A-intension. Suppose further that the criterion for 2-conceivability of a proposition is that there is at least one T in a given C-intension.[246] Propositions that are conceivable according to definition (CON) align with 2-conceivablility but

[246] If there is only one T in a certain C intension, this should occur when the actual and counterfactual world align (granted that it is impossible to have 2-conceivability without 1-conceivability).

not 1-conceivability, as (CON) entails being able to imagine the situation *determinately*, which 1-conceivability in isolation discounts. Further, propositions that are undecidable according to (b) are 1-conceivable but not 2-conceivable. In these cases, we can roughly conceive of a situation where a proposition is verified (that is, along the A-intension), but we cannot, even in principle imagine the situation determinately, because we do not know what the actual world is like. Propositions that are inconceivable according to (INC) are neither 1-conceivable nor 2-conceivable.

4.3 Assessment of "Zombies Are Conceivable"

When I imagine a philosophical zombie, do I imagine it *determinately*? If pressed, could I provide all the details about a philosophical zombie? If qualia is non-physical, all we would have to do is imagine the affirmation of being a physical duplicate and the negation of the non-physical phenomena constituting conscious experience, and provide every singular granular detail. That being said, providing every single detail is an impossible bar to set.[247] However, if we believe qualia is non-physical in the actual world, we are able to *in principle* imagine a philosophical zombie determinately and thus have 2-conceivability. Yet if qualia is physical, we cannot obtain 2-conceivability even in principle. Put simply, the 2-conceivability of qualia rests on the assumption that it is non-physical.

The 'conceivability' of zombies along the A intension betrays our lack of certainty in the physicalist thesis. If we discovered beyond doubt that the world is physicalist, then zombies would be inconceivable, as we would not view dualist worlds as genuine candidates for the actual world. Our *a priori*, or primary conceivability of zombies can be clarified with Yablo's account of undecidability,

[247] There are interesting arguments to be made regarding what details are relevant when conceiving of a situation. If the bar is set too low, critical details (such as if qualia is physical or non-physical) could be missed out, whereas if the bar is set too high, the criterion could end up being to detail an entire possible world. The latter seems to me a better bar to set if a correlation between conceivability and possibility is to be posited. Surely if all the relevant details about a situation can be articulated without inconsistency, then it is metaphysically possible. However, there may be seemingly unconnected details that, once articulated, actually produce a contradiction. Perhaps it is only once an entire possible world is articulated that we can be sure of the metaphysical possibility of a 'conceived' state of affairs. Yet at this point, we are positing something that looks more like an ersatz theory of possible worlds rather than conceivability.

interpreted according to (b). When we try to imagine a philosophical zombie from a basis of ignorance about which world is actual, what we imagine is by nature indeterminate. We cannot, even in principle, spell out the details of the world in which P&~Q is true, because we don't know what actual world our conceived situation is based on and therefore what we are negating in terms of Q. Therefore, our 1-conceivability of philosophical zombies can be explained by Braddon-Mitchell's account, that is, we are not one hundred percent certain which world is actual and therefore posit conceivability claims on this shaky foundation. Our purported 2-conceivability of zombies only holds on the assumption that dualism is true, which is a premise the conceivability argument aims to prove.

Chalmers' argument purports to prove that either materialism must be false, giving us 2-conceivability of zombies, or else Russellian Monism is true. However, our direct experience only gives us 1-conceivability, and 2-conceivability seems relegated to those who already have a firm dualist intuition.

5. Conclusion

We can now return to Braddon-Mitchell's original account for the zombie intuition in terms of two-dimensional semantics. In one sense, all conceivability and possibility claims are founded on a conception of what the world is actually like, meaning that P&~Q is conceivable for dualists and inconceivable for physicalists. Nevertheless, physicalists can acknowledge they may be mistaken, and thereby P&~Q is conceivable along the A-intension, in light of underlying uncertainty about what the actual world is like. Chalmers responds by accepting that conceivability along the A-intension does not entail metaphysical possibility, but that conceivability along the C-intension does, and further it can be deductively proven that the A and C intensions of P&~Q are identical. This argument at least partially relies on a sleight of hand that shifts from discussing physical objects with intrinsic profiles that differ to their structural profiles, to "microphysical properties" with differing intrinsic and structural profiles. The latter is incoherent. However, if we posit that microphysical properties have identical A and C intensions, then the same issue arises. That is, if structural relations alone do not necessitate qualia, then what are these intrinsic relations and what further work can they be doing?

If Chalmers' argument against materialism is accepted, this implies that we have secondary, or direct conceivability of zombies. This notion can be investigated with recourse to Yablo's work on conceivability and inconceivability according to (CON) and (INC). (CON) pertains to direct conceivability, and direct experience suggests we do not have direct conceivability of philosophical zombies, unless we already have a firm predilection to dualism. In the absence of direct conceivability, the primary conceivability of philosophical zombies remains intact, and can be interpreted according to Yablo's account as indeterminate, sitting between conceivability and inconceivability due to ambiguity surrounding the state of the actual world.

References

Braddon-Mitchell, David (2003) 'Qualia and Analytical Conditionals', *The Journal of Philosophy* **100**: 111–35. https://doi.org/10.5840/jphil2003100321.

Chalmers, David (2009) 'The Two‐Dimensional Argument Against Materialism', in A. Beckermann, B.P. McLaughlin, and S. Walter, eds. *The Oxford Handbook of Philosophy of Mind*: 313-336. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199262618.003.0019.

Jackson, Frank (1986) 'What Mary Didn't Know', *The Journal of Philosophy* **83**: 291-95. https://doi.org/10.2307/2026143.

Yablo, Stephen (1993) 'Is Conceivability a Guide to Possibility?', *Philosophy and Phenomenological Research* **53**: 1–42. https://doi.org/10.2307/2108052.

Epistemic Peerhood and the Epistemology of Disagreement

Chow Zhen Yi[248]

Nanyang Technological University

Abstract

The epistemic significance of peer disagreement plays a central role in social epistemology and affects many types of beliefs that we hold. For instance, religious, political and moral beliefs that are normally taken to be fairly personal and even sacred, are all potentially destabilised in the presence of peer disagreement. Conciliatory views on disagreement, in particular, argue that in the face of disagreement with someone you take as an epistemic equal, you are obligated to revise your doxastic attitudes towards the disputed belief. The purpose of this paper is to argue against Conciliatory views insofar as they assert that peer disagreements alone rationally require a person to revise their doxastic attitudes. My paper proceeds as follows: I offer an evaluation of the notion of epistemic peerhood, and conclude that utilising such notions of epistemic peerhood on Conciliatory views generates absurd results. I then propose that there should be other considerations on top of peer disagreement that should be taken into account for any doxastic revision to occur.

[248] Chow Zhen Yi is a final year undergraduate student at Nanyang Technological University. His philosophical interests also include areas of value theory (i.e., the nature of intrinsic value), the philosophy of well-being, and causation and responsibility in the philosophy of criminal law.

1. Introduction

> For I was conscious of knowing practically nothing.
>
> \- Socrates[249]

Peer disagreement is an interesting epistemic phenomenon where two individuals who are identified as epistemic peers (people with the same processing powers or epistemic background) disagree on the truth of some given proposition. Thomas Kelly, in illustrating peer disagreement, brings up the example of two people checking the temperature of a room:

'You and I are each attempting to determine the current temperature by consulting our own personal thermometers. In the past, the two thermometers have been equally reliable. At time t0, I consult my thermometer, find that it reads sixty-eight degrees, and so immediately take up the corresponding belief. Meanwhile, you consult your thermometer, find that it reads seventy-two degrees, and so immediately take up that belief. At time t1, you and I compare notes and discover that our thermometers have disagreed. How, if at all, should we revise our original opinions about the temperature in the light of this new information?'[250]

As asked by Kelly at the end of his illustration, how then should an epistemic agent respond appropriately to such peer disagreements? Different philosophers have concluded differently about the epistemic significance of peer disagreement, but there are two main camps set in response to this question: Conciliatory views and Steadfast view.[251] Given a disputed proposition between two epistemic peers, Conciliationists hold that the involved peers should modify their doxastic attitudes towards the disputed proposition such that both peers move their doxastic attitudes closer to the other party.[252] Proponents of the Steadfast view, on

[249] Plato. (2002) *Five dialogues Euthyphro, Apology, Crito, Meno, Phaedo* in Georges Maximilien Antoine Grube & John M Cooper, Trans., Hackett.

[250] Kelly, Thomas. (2010) 'Peer Disagreement and Higher Order Evidence', in Richard Feldman & Ted A Warfield, eds., *Disagreement*. essay, Oxford University Press.

[251] Matheson, Jonathan. (2015) *The epistemic significance of disagreement*. Palgrave Macmillan.

[252] See Feldman and Elga for Conciliatory views: Feldman, Richard. (2006) 'Epistemological Puzzles about Disagreement', in Stephen C Hetherington, ed., *Epistemology futures*. essay, Oxford University Press; Feldman, Richard. (2007) 'Reasonable Religious Disagreements', in Louise M. Antony, ed., *Philosophers without gods: Meditations on atheism and the secular life*, Oxford University Press; Elga, Adam. (2007) 'Reflection and disagreement' *Nous* **41**: 478–502.

the other hand, deny that there is a need to change one's doxastic attitudes towards the disputed proposition.[253] Rather, epistemic peers involved in a disagreement are entitled to hold on steadfastly to their prior beliefs even after occurrences of disagreement. In this paper, I will argue against Conciliatory views by arguing that peer disagreement alone is not a tenable basis for constituting defeaters for an epistemic agent's beliefs. In order to show this, I first argue that the general characterisation of epistemic peerhood cannot be used to assess a potential epistemic peer. I will then argue that proposed methods of assessing epistemic peerhood as articulated by philosophers in the disagreement literature, when utilised by Conciliatory views to generate peer disagreements, yields problematic consequences. These consequences include a form of extreme unqualified dogmatism such that our beliefs can never be wrong in the face of any disagreement, and an extreme form of epistemic scepticism that entails everyone knowing 'practically nothing'. I then propose that augmenting the conditions for peer disagreement to include propositional considerations on top of pure disagreements alone would help alleviate the issues raised.

The paper will proceed as follows: In section 2, I will elaborate on the general characterisation of epistemic peerhood. I then explain that there are two main ways of assessing whether an individual is an epistemic peer – (i) through the agreement and disagreement of auxiliary beliefs and (ii) through the evaluation of relevant credibility conferring features present in a potential epistemic peer. I will then explicate Conciliatory views further. In section 3, I apply (i) to Conciliatory views and show how it yields the peculiar result of justifying extreme unqualified dogmatism about our beliefs. In section 4, I apply (ii) to Conciliatory views and show how it renders any beliefs for any given epistemic agent to always be uncertain. Finally in section 5, I draw from the discussions in sections 3 and 4 to conclude that peer disagreement alone is insufficient to constitute a defeater for an epistemic agent's beliefs. I then propose that instead of revising doxastic

[253] See Kelly, Bergmann and Enoch for Steadfast views: Kelly, Thomas. (2010) 'Peer Disagreement and Higher Order Evidence', in Richard Feldman & Ted A Warfield, eds., *Disagreement*, Oxford University Press; Bergmann, Michael. (2015) 'Reasonable Religious Disagreements', in Jonathan Lee Kvanvig, ed., *Oxford Studies in Philosophy of Religion*, Oxford University Press; Enoch, David. (2010) 'Not just a truthometer: Taking oneself seriously (but not too seriously) in cases of peer disagreement' *Mind* **119**: 953–997.

attitudes based purely on instances of disagreement themselves, an epistemic agent should also consider relevant propositional content related to peer disagreements when making doxastic revisions.

2. Epistemic Peerhood and Conciliatory Views

Conciliatory views of peer disagreement heavily rely on certain characterisations of epistemic peerhood in order to work. In this section, I will first outline the core characteristics of epistemic peerhood as found in the disagreement literature before elaborating further on Conciliatory views. The general characterisation of epistemic peerhood is embedded within the wider peer disagreement framework, so it is to this framework first that we turn to. Nathan King proposes four conditions that accommodate most cases of disagreement within the peer disagreement literature:

1. *The disagreement condition*: S believes P, while T believes ~P
2. *The same evidence condition*: S and T have the same P-relevant evidence, E.
3. *The dispositional condition*: S and T are equally disposed to respond to E in an epistemically appropriate way.
4. *The acknowledgement condition*: S and T have good reason to think conditions (1)-(3) are satisfied.[254]

Conditions (1)-(4) cover two main areas central to generating genuine cases of peer disagreement. First, for an instance of disagreement itself, there has to be awareness on the part of both parties involved that each party holds opposing views (i.e., (1) S believes P while T believes ~P) and that both parties are aware that there is another party holding an opposing view (i.e., (4) the acknowledgement condition). Secondly, and vitally, both parties need to be epistemic peers, King here utilises (2) and (3) to cover the general idea of how an epistemic peer should be like. An epistemic peer would need to possess the same

[254] King, Nathan. (2011) 'Disagreement: What's the problem? or a good peer is hard to *find' Philosophy and Phenomenological Research* **85**: 252-253.

access to the same body of evidence in relation to a disputed claim (i.e., (2) the same evidence condition) and they also need to be equally able to process the aforementioned body of evidence (i.e., (3) the dispositional condition). Analogously, Lougheed also notes that common standards used by many philosophers to characterise epistemic peerhood are, in fact, *cognitive* and *evidential* standards.[255] This aligns well with King's conditions of (2) and (3), such that epistemic peers need to have the same cognitive abilities insofar as they may possess the same epistemic virtues like reliability, honesty or having good memory, or have the same cognitive processing power whilst evaluating relevant evidence (i.e., following (3)). Epistemic peers also need to possess the same body of evidence (i.e., following (2)). King and Lougheed's survey of the general character of epistemic peerhood then yields two defining characteristics: the same evidential possession and evidential processing abilities.[256]

Despite a rough consensus on what an ideal epistemic peer looks like, these characteristics are admittedly difficult to detect in an epistemic agent in real-world scenarios. After all, outside of idealised disagreements or what Elga calls 'clean'

[255] These citations are from Lougheed: Lougheed, Kirk. (2020) *The epistemic benefits of disagreement*. Springer; Lackey, A Jennifer. (2014) 'Taking Religious Disagreement Seriously', in Laura Frances Callahan & Timothy O'Connor, eds., *Religious faith and intellectual virtue*, Oxford University Press; Oppy, Graham. (2010) 'Disagreement' *International Journal for Philosophy of Religion* **68**: 183–199.

[256] These two characteristics are also mentioned by Kelly, Thomas. (2005) 'The Epistemic Significance of Disagreement', in Tamar Z Gendler & John Hawthorne, eds., *Oxford studies in epistemology*, Clarendon Press; Matheson, Jonathan. (2015) *The epistemic significance of disagreement*. Palgrave Macmillan; Elgin, Catherine. (2010) 'Persistent Disagreement', in Robert Feldman & Ted A Warfield, eds., *Disagreement*, Oxford University Press. It however should be noted that many other philosophers require *equal* evidence or processing abilities, it does not matter that the relevant bodies of evidence possessed by two epistemic agents are different or that each have different epistemic virtues, as long as both bodies of evidences or overall cognitive abilities are *equally* good both epistemic agents are epistemic peers (See footnote 19 for more details). This is distinguished from the exact *same* evidence or processing abilities. In other words, the type of epistemic virtues, the intellectual background and the relevant body of evidence needs to be close to identical. Nevertheless, for the sake of argument I utilise *equal* conditions instead of *same* conditions, this is as I believe same conditions to be a proper subset of equal conditions. After all, if two conditions are identical (the same), they necessarily have to be equal as well. Additionally, Choo agrees with using equal instead of same conditions, he argues that 'this is because such an understanding of peerhood seems to be the driving force behind the different principles and arguments in the literature. For example, conciliationists think that one should give equal weight to one's disputant in cases where both possess the same evidence and same dispositions *because* having the same evidence and dispositions make both equally likely to be correct in that scenario': Choo, Frederick. (2018) 'The epistemic significance of religious disagreements: Cases of unconfirmed superiority disagreements' *Topoi* **40**: 1139–1147

or 'pure' cases of disagreement,[257] it is practically impossible to accurately determine whether a potential epistemic peer possesses the exact same body of evidence as oneself, or if they are able to process relevant pieces of evidence at an equal proficiency as well. Even close to ideal cases of peer disagreement would not seem to suffice to allow one to accurately establish if someone is an epistemic peer. Suppose that there are two philosophy professors who disagree on some philosophical proposition, both possess the same credentials from the same university and more so work as professors in the same philosophy department. Further, both have the same number of publications in the same specialised field. Yet, even in such cases, it seems difficult to assert that both professors would possess the exact same or equal cognitive background or that both are aware of the exact same or equal body of evidence. After all, even with such circumstances in place, either professor could nonetheless have differing overall epistemic virtues or fail to have read a certain relevant publication that counts as evidence. As King notes, such cognitive standards or dispositional conditions require equality of 'reliability with respect to the relevant field of inquiry' where 'general intelligence and logical skill' must also be equal. Furthermore, 'intellectual virtues', 'epistemic virtues such as honesty, carefulness and freedom from bias', and similarity of 'background beliefs' also need to be equal.[258] The complexity of the necessary and sufficient sub-conditions of dispositional conditions/cognitive standards inevitably make it difficult, even in close to ideal real-world cases like that of the two philosophy professors, to determine who an epistemic peer is. King further notes that this complexity also extends to the characteristic of equal evidential possession. He asserts that 'for many disagreements in philosophy and in other fields … intelligent, similarly trained subjects possess bodies of evidence that are overlapping but not co-extensive',[259] and this only considers arguments as evidence. King goes on to note that evidence may also include 'such items as perceptual experiences, rational insights, 'seemings', or intuitions' which influences a subject's doxastic attitudes in ways that cannot be fully

[257] Elga 'Reflection and disagreement', 492.

[258] King 'Disagreement: What's the problem? or a good peer is hard to find', 257-259.

[259] King 'Disagreement: What's the problem? or a good peer is hard to find', 255.

communicated to other people in a propositional sense.[260] Thus, given the epistemic limitations of individual people and the fact that most people do not share coextensive bodies of evidences, the difficulty in assessing epistemic peers based on the characterisation of epistemic peerhood itself is thus further worsened.

How then, should an epistemic agent go about assessing a potential epistemic peer? Frederick Choo suggests that there are two main ways of assessing whether someone is or is not an epistemic peer:[261] First is a method articulated by Adam Elga, here Elga proposes that for 'messy real-world cases', the criterion necessary to assess whether a person is an epistemic peer consists of whether there is a wide-ranging agreement or disagreement of related claims to the disputed claim.[262] There is thus a sort of symmetry between epistemic peers: epistemic peers would have sets of beliefs that are largely in agreement with each other and non-peers would not. In order to illustrate this, he brings up the example of Ann and Beth who are disagreeing on the issue of abortion. For Elga, Ann and Beth have discussed closely related auxiliary claims like whether 'human beings have souls', or whether 'it is permissible to withhold treatment from certain terminally ill infants' and both disagree on all these auxiliary claims.[263] From Ann's perspective, given that Beth is wrong on these many related auxiliary claims, this gives Ann reason to dismiss Beth as an epistemic peer in relation to the disputed claim about abortion. Conversely, if there is wide ranging agreement on related auxiliary claims, Ann has reason to see Beth as an epistemic peer. Thus, Elga argues that 'with respect to many controversial issues, the associates who one counts as peers tend to have views that are similar to one's own.' and epistemic peers are those 'who agree with you on issues closely linked to the one in question.'.[264]

[260] King 'Disagreement: What's the problem? or a good peer is hard to find', 256.

[261] Choo, 'The epistemic significance of religious disagreements: Cases of unconfirmed superiority disagreements'.

[262] Elga, 'Reflection and disagreement' 492.

[263] Elga, 'Reflection and disagreement' 493.

[264] Elga, 'Reflection and disagreement', 494.

The second method as suggested by Choo, is a relatively more straightforward method of basing epistemic peerhood on 'relevant credibility-conferring features (i.e. evidential possession and processing)'.[265] Choo does anticipate the difficulties raised above about establishing peerhood based on the general characterisation of evidential possession and processing. In response, he suggests that a potential approach to avoid such difficulties is to 'look at one's track record'.[266] In elaboration he states:

> A track record may consist of various things like testimony, institutional certification, having been right in previous disagreements, and so forth. This may tell us if our disputant has the relevant credibility-conferring features without having to identify and assess the specific features.[267]

Choo here adopts a more inductive or probabilistic approach in assessing epistemic peerhood. The level of evidential possession and processing an epistemic agent possesses is in some inductive sense antecedent to the type of track record they have. Hence, by analysing a potential epistemic peer's track record, one can get a possible picture on what their evidential possession and processing capabilities are like. Thus, if someone has a good track record that matches my own, I can take it that there is a good probability that they are my epistemic peer.

[265] Choo, 'The epistemic significance of religious disagreements: Cases of unconfirmed superiority disagreements', 1141. It should be noted that many philosophers working in the epistemology of disagreement generally are in agreement with notion that relevant credibility-conferring features are vital in assessing a potential epistemic peer. Take for instance Thomas Kelly who argues that individuals are epistemic peers insofar as '(i) they are equals with respect to their familiarity with the evidence and arguments which bear on that question' and '(ii) they are equals with respect to general epistemic virtues such as intelligence, thoughtfulness and freedom from bias': Kelly 'The Epistemic Significance of Disagreement',174. Richard Feldman, for example, in identifying epistemic peers asserts that 'people are epistemic peers when they are roughly equal with respect to intelligence, reasoning powers, background information, etc.': Feldman 'Reasonable Religious Disagreements', 144. Jonathan Matheson thinks if someone is an epistemic peer they will have 'distinct, but equally good, bodies of evidence' and that 'the likelihood of their processing the evidence correctly is equally high': Matheson 'The epistemic significance of disagreement' 22.

[266] Choo, 'The epistemic significance of religious disagreements: Cases of unconfirmed superiority disagreements', 1141.

[267] Choo, 'The epistemic significance of religious disagreements: Cases of unconfirmed superiority disagreements', 1141-1142.

Generally, Conciliatory views assert two theses: (i) When there is a peer disagreement between two epistemic peers A and B about an issue P, the awareness of the peer disagreement alone constitutes a sort of defeater for both A's and B's belief about P prior to the disagreement. And (ii) the defeater generated from the disagreement requires both epistemic peers A and B to modify or change their doxastic attitudes regarding P.[268] Regarding (i), as can be noted from the above discussion, when someone is determined to be an epistemic peer, they would be seen as a reliable indicator of the correct interpretation or inference of some given proposition. Thus, when made aware of a reliable indicator of a proposition that opposes one's own views, this should be taken as a reason (or a defeater) that one's own views may be wrong. This then motivates (ii) such that the involved epistemic peers in the disagreement are now rationally obligated to modify their doxastic attitudes to be appropriately aligned with the new available reasons on what the correct interpretation of a disputed proposition is.

Given the overview on the characterisation and assessment of epistemic peerhood, and how Conciliatory views generally work, we can then inquire about the main contention of this paper: How does the above-mentioned assessment methods of epistemic peerhood work against Conciliatory views of disagreement? Or more precisely, why is disagreement between epistemic peers untenable as a basis for constituting defeaters for disputed beliefs (i.e., Conciliationism)? It is to these arguments that we now turn to.

3. Elga's Wide Ranging Agreement/Disagreement and Conciliatory views

As mentioned, Elga's method of peer assessment relies on wide ranging agreement or disagreement on related claims or beliefs. Despite this method's simplification of the peer assessment process, it nevertheless runs into quite a huge problem when applied to Conciliatory views. Jennifer Lackey in particular points out that:

[268] Matheson 'The epistemic significance of disagreement'.

> A … problem with Elga's view here is that while he advertises it as 'conciliatory' where 'equal weight' is given to one's own belief and to that of one's opponent, it sanctions a dogmatic 'sticking to one's guns' in nearly all of the cases of disagreement that are of deep importance to us. Disagreements regarding religious, moral, political, and philosophical matters, for instance, almost invariably involve opposing views about a range of related issues that will lead the relevant parties to fail to count one another as epistemic peers. On Elga's view, then no doxastic revision is required in all of these cases. Not only is this a peculiar result for a 'conciliatory' view, it also seems epistemically wrong – surely there are some cases where at least some doxastic revision is rationally required when disagreeing about contentious matters, even when the disagreement involves a host of related questions.[269]

I believe Lackey rightly points out that Elga's views entail that for any given controversial disagreement (i.e., on religion, morality or politics), opposing parties would fail to recognise each other as epistemic peers in the first place. Unfortunately, Lackey does not elaborate further on how exactly does disagreements on a certain issue 'invariably involve opposing views about a range of related issues.'[270] In other words, there needs to be an argument elaborating exactly how a disagreement on a disputed issue would necessarily involve disagreement on many related issues as well. I propose that Elga's assessment method entails such consequences due to a presupposition of some coherence relation that holds between a disputed belief and auxiliary related beliefs. I will take it a step further than Lackey and argue that Elga's views not only leads to relevant parties failing to see each other as epistemic peers in controversial disagreements, but also in *all* disagreements. If such an argument is successful, it would demonstrate that Elga's method of assessment practically renders no one to be an epistemic peer, or that sticking to conciliatory views with Elga's assessment methods allows unqualified dogmatism.

[269] This citation from Lackey is from Lougheed: Lougheed, 'The epistemic benefits of disagreement'; Lackey 'Taking Religious Disagreement Seriously' 308.

[270] Lackey 'Taking Religious Disagreement Seriously' 308.

As mentioned above, Elga uses the terms 'closely linked' or 'closely related' to describe the relation between disputed and auxiliary claims. Nevertheless, Elga does not clearly specify what it actually means for auxiliary claims and a disputed claim to be 'closely related' (insofar as the relation between both types of claims makes the auxiliary claims relevant for the evaluation of someone as an epistemic peer). Thus, in order to try and explicate how Elga understands the relationship between disputed and auxiliary claims, we need to turn to the examples that he raises. In elaboration of his criterion of wide-ranging agreement/disagreement, Elga uses two examples: the Ann and Beth example as noted above and the political framework example. Let's look at the Ann and Beth example first.

In the Ann and Beth example, the disputed claim here is whether abortion is morally permissible. Elga notes that 'claims *closely linked* to the abortion claim' include 'whether human beings have souls, whether it is permissible to withhold treatment from terminally ill infants, and whether rights figure prominently in a correct ethical theory.'[271] The question then arises, what exactly makes these claims closely linked to the abortion claim? Why would a certain stance taken on whether human beings have souls relate to a stance taken on whether abortion is morally permissible? The answer here seems to be that they have inferential and/or explanatory connection: a stance taken on one issue can help an epistemic agent *infer* and *explain* another stance taken on another issue.[272] For instance, if Ann holds the stance that human beings do indeed have souls, and, perhaps, assuming she also believes that human foetuses are endowed with souls, it is not hard to see why in Ann's eyes, she *infers* that abortion would be like murder. This would also allow Ann to *explain* why abortion is morally impermissible: it would be the ending of a human life.

[271] Elga, 'Reflection and disagreement'493, emphasis mine.

[272] Of course, one can argue that the connections between auxiliary and disputed claims are not explanatory in nature. But it is hard to imagine that they are not, they seem to have some kind of semantic relationship where the meaning of a stance taken on one claim influences the meaning of the stance taken on another claim. What does humans having souls have to do with the moral permissibility of abortion? Clearly the former claim serves as a premise of sorts in understanding the answer to the latter. An objector would have to provide an argument as to why this intuitive explanatory connection is not the case between stances taken on auxiliary and disputed claims.

Elga's political framework example seems to parallel the above sentiment. He writes that when 'a smart and well-informed friend who has a basic political framework diametrically opposed to your [the reader's] own' runs into a disputable new political claim *y*, the disagreement on the auxiliary claims (the opposed basic political framework) should allow the reader to infer that 'you [the reader] are more likely than your friend to correctly judge'[273] the new political claim *y*. How exactly does one's basic political framework affect the correctness of one's stance on the new political claim *y* in the eyes of a disputant? Again, through explanatory and/or inferential connections: if you think the basic political framework of your friend is wrong, then their stance on a disputed claim inferred from their wrongheaded basic political framework would also be wrong.

How does this relate to the coherence relation mentioned above? As it stands, explanatory connections and inferential connectedness are essential ingredients in a coherence relation. According to Noah Lemos, most epistemologists often cite at least three factors in characterising coherence relations: (i) logical consistency, (ii) explanatory connections, and (iii) conformity with norms about belief formation.[274] Laurence BonJour, a major defendant of coherentism, also echoes Lemos' assertions. BonJour argues that a coherence relation should at least contain more than logical consistency, have explanatory connections, and inferential connectedness.[275] For our purposes, only (i) and (ii) are directly relevant.[276] For (i), avoiding logical inconsistency would support the explanatory or inferential

[273] Elga, 'Reflection and disagreement'493.

[274] Lemos, Noah M. (2021) *An introduction to the theory of knowledge*. Cambridge University Press, 73.

[275] Bonjour takes explanatory connections to be a special species of inferential connections. He holds that not all inferential connections explain things, an explanatory connection not only allows inferences to be made but it also does not leave unexplained anomalies in a set of beliefs: BonJour, Laurence. (1985) *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press,

[276] The reason I take (iii) to not be directly relevant is that we are trying to see how Elga's wide-ranging criterion assumes some sort of coherence relation, and (iii) does not help us see if Elga's criterion resembles a coherence relation. It is true that beliefs that are formed through epistemic vices, like beliefs formed dishonestly or through wishful thinking, would seem less coherent; However, the flouting of such epistemic norms like honesty or realistic thinking (if that is an epistemic norm) does not clearly impact an agent's agreement or disagreement on auxiliary issues with a potential peer – which is the crux of this section. As well it will be further discussed, (i) and (ii) does indeed directly impact an agent's agreement and disagreement on auxiliary issues with a potential peer.

connectedness of auxiliary and disputed claims.[277] After all, logical inconsistency negates any possible explanatory or inferential connections between claims.[278] For (ii), as illustrated above, the relationship between stances taken on auxiliary and disputed claims seems well characterised as explanatory and/or inferential connections. Given this, the relation between stances taken on auxiliary and disputed beliefs would seem to require (i) and (ii), this however would minimally make it some type of coherence relation.

The issue is, if the relation between a stance taken on auxiliary beliefs and disputed beliefs is one of coherence, such a relation would be bi-directional in the sense that it would allow inferences from stances on auxiliary claims to stances on disputed claims *and vice versa*. To elaborate, if a disputant P holds a stance X_1 about disputed claim a, then P has to hold stance X_1 about related auxiliary claims b, c, and d as well. However, the opposite is also true: if P holds a stance X_1 about related auxiliary claims b, c, and d, then P has to hold stance X_1 about disputed claim a. This is as the coherence relation would require P to avoid logical inconsistency (or hold onto logical consistency)[279] for the sake of the explanatory connections between a, b, c, and d. Notice, however, if another disputant Q holds some stance X_2 (that is inconsistent with X_1) about a, the coherence relation would require Q hold X_2 about b, c, and d as well. This would entail that given some disputed claim a, where P holds X_1 and Q holds X_2 about a, P and Q would also have to hold X_1 and X_2 respectively for all related auxiliary claims b, c, and d.[280]

[277] BonJour holds that not only should there be no logical inconsistency/logical consistency, there should be probabilistic consistency as well – such that two claims have greater coherence insofar as either makes the other far more probable. He puts the conditions as such '(1) A system of beliefs is coherent only if it is logically consistent. (2) A system of beliefs is coherent in proportion to its degree of probabilistic consistency.': BonJour 'The structure of empirical knowledge', 95.

[278] For instance, consider the Ann and Beth example again. Imagine that Ann holds a stance on auxiliary claims (i.e., humans have no souls or no rights figure in any ethical theory) that amounts to or entails the stance that abortion is in fact morally permissible, it would be hard to see how Ann's stance on auxiliary claims provides any explanatory or inferential support for the inconsistent stance that abortion is in fact *not* morally permissible.

[279] The coherence relation could also require probabilistic consistency (see footnote 33) instead. In any case, this switch does not change my argument. What I am trying to establish is that the coherence relation between stances taken on auxiliary and disputed claims provides good reason for an agent P to maintain the same stance X_1 for both auxiliary and disputed claims.

[280] For clarity's sake, consider a question like 'what kind of existence do numbers have?' an eliminative physicalist may answer that numbers have concrete existence while a Platonist may answer that numbers have an abstract existence. It would very well be fair for the physicalist to

However, given the above argument, Elga's criterion would entail that for any given disagreement about some disputed claim a where disputants P and Q have differing stances on a, P and Q would then have good reason to infer that either party would also disagree about a wide range of related auxiliary claims to a. Notice, however, that this would have the unsavoury consequence of allowing any epistemic agent involved in any disagreement to dismiss any person disagreeing with them as an epistemic peer. This, as Lackey notes, 'is … a peculiar result for a 'conciliatory' view' and that such a consequence 'seems epistemically wrong' insofar as it justifies a sort of extreme unqualified dogmatism that does not require any doxastic revisions to our beliefs even in the face of disagreements.[281] In fact, such a method of assessing epistemic peerhood leaves Conciliatory views practically irrelevant, as given Elga's criterion, there would be no genuine cases of peer disagreement generated at all. Nevertheless, there is a possible objection lurking. Notice that we have been discussing peer disagreement in a normative sense, such that we have been assuming that epistemic agents would act rationally in accordance with certain rational norms (i.e., consistency). It can, however, be argued that human behaviour regarding their beliefs is not at all aligned to what the coherence relation (i.e., consistency) rationally requires them to do. In reality, humans are irrational, humans are not descriptively beholden to consistency or things like the coherence relation where if one belief is wrong, it is likely a related held belief is also wrong. Human beings frequently hold contradictory beliefs regarding a wide range of related issues; People actually or realistically use beliefs that they do have (i.e., inconsistent beliefs) rather than beliefs that they should have (i.e., consistent beliefs) to discern epistemic peers. Thus, it can be the case that people may be right about one issue, but are wrong about many related issues.

In this sense, Elga's criterion of similar auxiliary beliefs can be interpreted descriptively, such that when people discern out epistemic peers, it just so

infer from the Platonist's answer that the Platonist likely believes that physicalist positions on other philosophical areas are false – this is as strong or eliminative physicalism is committed to only items of concrete existence and Platonism is committed to the existence of abstract objects. This is what I mean when I say that the stance taken on auxiliary claims relating in an explanatory way to a disputed claim *explains* why a certain stance is taken on a disputed claim.

[281] Lackey 'Taking Religious Disagreement Seriously', 308.

happens to be the case that there is a positive correlation between someone being a peer (having many agreed upon related claims) and agreement on a disputed claim. There is no coherence relation involved, and consequently even when there is a disagreement between peers regarding a certain issue, it does not mean that either has to be wrong on related auxiliary issues as well. How can we respond to this objection? As mentioned above, we have been discussing peer disagreements normatively, this however is because Conciliatory views themselves are normative insofar as they obligate epistemic agents to *do* something about their doxastic attitudes. Conciliatory views when applied to the real world still seem to require an acceptance of logical consistency about beliefs in order for them to work. This is as Conciliationism holds that when there are two conflicting stances on a single issue, both cannot be completely right as they are inconsistent. Consequently, due to this inconsistency, epistemic peers in conflict need to revise their doxastic attitudes towards their own respective beliefs. However, if logical consistency is unnecessary in real life, then Conciliationism would fail to have any obligatory strength to motivate epistemic agents to revise their doxastic attitudes in the first place.

Thus, given that (a) the relationship between auxiliary beliefs and the disputed belief is one of coherence (which includes consistency), and that (b) under Conciliatory views, human beings have to be beholden to the epistemic virtues of consistency or coherence etc. when it comes to holding a set of beliefs (even in real life). Both (a) and (b) together entail that given any disagreement between two people, even if both people seem to be epistemic peers through *prima facie* shared auxiliary beliefs, both parties are also justified in thinking that if the other party comes to a different stance on some disputed claim, they must have gotten something wrong about their stances on auxiliary claims due to the logical consistency of their set of beliefs (and thus cannot be epistemic peers even if it *prima facie* seemed to be the case originally). But, since Elga's criterion of epistemic peerhood seems to assume (a) and Conciliationism requires (b), this entails that any disagreement over a disputed belief would still allow involved epistemic agents to justifiably dismiss the disagreeing party as an epistemic peer.

4. Credibility-Conferring features, Track Records and Conciliatory views

Recall that the second method of assessment discussed in section 2, as elaborated on by Choo, involves looking at a potential epistemic peer's track record to get a significant enough sensing of their relevant credibility-conferring features. This seems like a promising enough method, yet, this method too runs into some serious difficulties. In order to understand this difficulty, we first have to see why there seems to be no need for any presently existing actual peers for peer disagreements to possess any epistemic significance. Kirk Lougheed brings up the following thought experiment as an illustration:

> Suppose that Peter van Inwagen and David Lewis are the only two experts on evidence for compatibility of free will and determinism. They alone are epistemic peers with each other, at least with respect to the question of whether free will and determinism are compatible. Imagine that they are both flying together to a conference where they will give competing presentations on whether compatibilism is true. But the plane malfunctions and crashes into the ocean. Lewis is the only survivor and manages to swim to a small island. Now that van Inwagen has perished Lewis has no actual peer with respect to compatibilism … Lewis was aware that only van Inwagen was his peer and they disagreed about compatibilism. Surely, Lewis cannot reasonably dismiss the significance of peer disagreement simply by pointing out van Inwagen does not exist anymore. Therefore, the epistemic significance of peer disagreement, whatever it may be, does not require that an actual peer presently exists.[282]

Lougheed's thought experiment does satisfy King's four conditions for genuine peer disagreement as discussed above. Yet, it does have the implication that there is no need for any actual peers in order for peer disagreements to have epistemic significance. Thus, under Conciliatory views, epistemic agents may need to revise their doxastic attitudes even if there is no actual peer present that disagrees with them. How is this a problem for the track record method? Consider first, as Nathan Ballantye points out, that there could be counterfactual versions of us that

[282] Lougheed 'The epistemic benefits of disagreement', 42.

arguably know far more about a subject than our actual selves.[283] Take philosophical disagreements as an example, Ballantye argues that 'we know that we regrettably do not have all the arguments, distinctions, and objections that the counterfactual philosophers would have devised.'[284] Lougheed further agrees with Ballantye here, asserting that this shows that 'counterfactual peer disagreement is just as epistemically significant as actual peer disagreement.'[285] Consequently, it is quite plausible to take counterfactual versions of ourselves to minimally have a similar total epistemic position as compared to the actual versions of ourselves. In terms of the track record assessment method, it is easy to imagine a counterfactual version of us in a nearby possible world that, while possessing a similar track record as compared to the actual us, holds onto differing relevant opinions. As mentioned earlier, Choo suggests that 'a track record may consist of various things like testimony, institutional certification' or 'having been right in previous disagreements.'[286] Consider me as an example, I currently am working towards a bachelor's degree in philosophy. Perhaps, my friends and professors think I have some decent ability in philosophy, and this has been demonstrated in previous classes and conversations. Here, I would have a certain level of institutional certification, testimony from friends and professors, and having instances of being right (sometimes) in previous disagreements in class etc. Nevertheless, it is clearly logically possible that there is a counterfactual version of me that has attended another university also majoring in philosophy, taken similar modules as the actual me, my friends and professors also think I am decent in philosophy, and I also have proven so in class. Yet, it is entirely logically possible that counterfactual me holds differing opinions on many philosophical

[283] Ballantyne, Nathan. (2014) 'Counterfactual Philosophers' *Philosophy and Phenomenological Research* **88**: 368–87.

[284] This citation is from Ballantye is from Lougheed: Lougheed 'The epistemic benefits of disagreement'; Balllantye 'Counterfactual Philosophers' 368.

[285] Lougheed 'The epistemic benefits of disagreement',43.

[286] Choo, 'The epistemic significance of religious disagreements: Cases of unconfirmed superiority disagreements', 1141.

topics as compared to the actual me.[287] If Lougheed and Ballantye are right and 'counterfactual peer disagreement is just as epistemically significant as actual peer disagreement',[288] then for every philosophical belief that I have, I am required under Conciliatory views to constantly revise my doxastic attitudes to be less certain about my beliefs. This is as there is always going to be a counterfactual version of myself with a similar track record but opposing beliefs about any given philosophical issue. Crucially, this results in an untenable skepticism about my philosophical beliefs, as for every new belief P I might take on or modify, I will have to immediately revise my doxastic attitude towards P. I would consequently end up 'knowing practically nothing.'

This counterfactual conundrum that the track record method faces, strikingly extends beyond just philosophical beliefs or counterfactual versions of ourselves. It would seem that if Conciliatory views and counterfactual peer disagreement are both acceptable, then this would fundamentally destabilise all our beliefs. Given that for any belief that we might hold, there can be a counterfactual peer (whether ourselves or someone else) that has a similar track record and opposing views. This would entail that for any new belief Q pertaining to literally anything, under the track record method, I would have to yet again constantly revise my doxastic attitudes towards Q, or any new beliefs I take on to replace Q. Again, an untenable skepticism arises, this time pertaining to all our beliefs.

5. Further significance of peer disagreement on doxastic attitudes

I have argued above that both the methods of assessing epistemic peerhood are indeed quite problematic when used alone to establish the epistemic significance of peer disagreement (i.e., Conciliatory views). Given this difficulty, it can then be

[287] Due to some feedback from an anonymous reviewer, I should further elaborate/clarify this point. It is possible to conceive of a counterfactual version of me who despite having a similar track record – say going to the exact same university as actual me, demonstrating a similar level of philosophical skills, taking the same ethics classes as actual me – comes to a plausible view that is different from my own. Say, I hold abortion to be morally permissible, yet counterfactual me despite having a similar track record holds that abortion is not morally permissible. If counterfactual me is considered a peer, this would under conciliationist views, require me to revise my doxastic attitudes towards my belief that abortion is morally permissible.

[288] Lougheed 'The epistemic benefits of disagreement', 43.

asked, what should the epistemic significance of peer disagreements be like for epistemic agents? Recall that Conciliatory views hold that the presence of a genuine peer disagreement *alone* constitutes a sort of defeater for disputed beliefs that interlocutors in the disagreement.[289] I think in answering the previous question, a broader approach that is more sensitive to surrounding contextual factors of a given peer disagreement is necessary. Alvin Plantinga in discussing the nature of defeaters asserts that:

> Defeaters depend on and are relative to the rest of your noetic structure, the rest of what you know and believe. Whether a belief *A* is a defeater for a belief *B* doesn't depend merely on my current experience; it also depends on what other beliefs I have, how firmly I hold them, and the like.[290]

Plantinga's point here can be contextualised with an example of his. Consider the famous mathematician and philosopher Gottlob Frege. As Plantinga writes, Frege once believed that:

> (F) For every condition or property *P*, there exists the set of just those things that have *P*.[291]

Bertrand Russell then famously wrote Frege a letter pointing out serious issues in (F), one of Russell's famous paradoxes, that show how (F) if true, generates consequences that prove (F) itself to be false. As Plantinga notes 'before he realised this problem with (F), Frege did not have a defeater for it. Once he understood Russell's letter, however, he did; and the defeater was just the fact that (F) … entails a contradiction.'[292] Here it can be assumed that both Russell and Frege are roughly epistemic peers, both are famous mathematicians/philosophers

[289] There are generally two types of defeaters – rebutting defeaters and undercutting defeaters. For rebutting defeaters, these are defeaters that rebut a belief in some proposition *p*, such that it shows *p* to be false and therefore rationally inconsistent to continue holding onto *p*. Undercutting defeaters, on the other hand, undercut your justification or grounds for believing in *p*, thus instead of showing *p* to be false, undercutting defeaters show that there is no reason to take *p* to be true. Depending on the specific Conciliatory views peer disagreement can generate defeaters of either kind: Matheson' The epistemic significance of disagreement'; Plantinga, Alvin. (2000) *Warranted Christian belief*. Oxford University Press; Bergmann, Michael. (1997) 'Internalism, externalism and epistemic defeat' (dissertation), University of Notre Dame.

[290] Plantinga 'Warranted Christian Belief' 360.

[291] Plantinga 'Warranted Christian Belief' 361.

[292] Plantinga 'Warranted Christian Belief'.

working in the same era and in the same field of mathematics and logic. Yet, applying Plantinga's point here, it is not the mere peer disagreement of the truth of (F) between Russell and Frege that caused Frege to change his mind about (F), rather it is the propositional content of the disagreement between Frege and Russell. It is the argument or reasons provided by Rusell as an epistemic peer that ultimately provided the impetus for Frege's change of mind about (F). In other words, Frege's change of mind is due to the sort of propositional interaction between his previous beliefs in his noetic structure about (F) and the arguments or reasons provided by Russell against (F).[293]

What I am suggesting here is that there should be other conditions on top of just peer disagreement alone that constitutes a defeater for a person's belief. From section 3 and 4, we can see that if epistemic peerhood alone is both the necessary and sufficient conditions for disagreements to have epistemic significance, deeply problematic issues can arise. Therefore, I propose that at least another necessary condition for generating genuine defeaters should be paired alongside epistemic peerhood, namely, the propositional content related to the disagreement should be given central focus. Epistemic peerhood provides a marker of credibility insofar that the propositional content of a potential disagreement with a peer should be

[293] A potential issue here is that, one could point out that if Frege was unaware of Russell's paradox, then Frege and Russell would not have possessed the same evidence anyway. Thus, Frege and Russell were not epistemic peers to begin with. Two things can be said here. First, as discussed in section 2, the condition of same/equal evidential possession is incredibly difficulty to determine practically, and is therefore quite untenable in determining who an epistemic peer is. Therefore, if we relegate to assessing epistemic peers with the methods as suggested by Choo, we can still conclude that Frege and Russell should roughly be considered as epistemic peers. Secondly, even if we concede that for someone to be identified as an epistemic peer they must have the same evidence base, recall from footnote 8 that the condition for *equal* evidential possession is preferred over the condition for the *same* evidential possession. Frege and Russell could have very well *equal* evidence bases even though they do not have the *same* arguments. Therefore, even if Frege did not possess Russell's argument against (F), he could have possessed arguably equally good arguments for (F). Nevertheless, it is only when Frege became aware of the propositional content of Russell's argument against (F) which (let's assume) is as good as his own arguments for (F), that Frege chose to revise his doxastic attitudes toward (F). Notice, that the basis for this doxastic revision necessarily includes a potential defeater's propositional interaction with a person's noetic structure. It is not based on pure disagreement alone, but instead on the propositional interaction between Frege and Russell's evidence bases. In the same way, Russell could have possibly chosen to revise his doxastic attitude towards (F) instead, given some argument Frege could have made. For instance, Russell's argument has been rejected propositionally by many Neo-Fregans: Hale, Bob, and Crispin Wright. (2001) *The reason's proper study: Essays towards a neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon.

taken seriously. For example, if a young child, a friend and a physics professor tell me a plate can float, I am inclined to not take the reasons the young child offers seriously, as I do not count them as a peer. But, if it were instead my friend or the professor (epistemic superior), I would be inclined to listen carefully to the reasons they give and ultimately base any doxastic revision on those reasons offered. If propositional content is taken to be as the central focus of the epistemic significance of peer disagreements, the problems raised in section 3 and 4 either evaporate or are mitigated. Let us consider each, for section 3, the main issue being extreme unqualfied dogmatism. If propositional content is allowed to contribute to the epistemic significance of peer disagreement, then for any possible disagreement, it is first and foremost the propositional content of the reasons given by an individual that should have an impact on my doxastic attitudes. If a child tells me she sees someone dangerous in the bathroom even though I did not notice anyone around, I still take what she says with a level of doxastic weight (even if she is not my peer), and thus change my doxastic attitudes towards the idea that there might be someone in the bathroom. It is the propositional content of that claim that seems entirely coherent (given interaction with my noetic structure) and consequently possible, therefore I would decide to entertain the claim. I might not put as much doxastic weight onto a child's testimony as compared to a peer, but nevertheless, I still do due to the testimony's propositional content's relative consistency with my noetic structure. As for section 4, counterfactual peer disagreement lacks any actual propositional content, for we would not know what exact arguments a counterfactual peer would have for disagreeing with us. Thus, with propositional content, counterfactual peer disagreements would no longer yield an untenable scepticism.

6. Conclusion

I have argued that both characterisations and assessments of epistemic peerhood run into serious epistemic difficulties when used in Conciliatory views, suggesting that peer disagreement alone is insufficient for contributing significant epistemic significance for any doxastic revisions. Of course, there are other possible accounts of epistemic peerhood that are not covered in this paper, but I believe that I have covered most accounts of peerhood in at least broad strokes. Thus, my arguments hopefully have a wide resonance against most Conciliatory views.[294]

[294] I would like to thank Prof Grace Boey for her feedback on an early draft of this paper and for her invaluable class on social epistemology - without which, this paper would not have been possible. I would also like to thank Frederick Lim, Benjamin Tan, Ang Tse Pin and Allysa Escanuela for their valuable input and discussions.

References

Ballantyne, Nathan. (2014) 'Counterfactual Philosophers' *Philosophy and Phenomenological Research* **88**: 368–87. https://doi.org/10.1111/phpr.12068.

Bergmann, Michael. (1997) *Internalism, externalism and epistemic defeat* (dissertation), University of Notre Dame.

Bergmann, Michael. (2015) 'Reasonable Religious Disagreements', in Jonathan Lee Kvanvig, ed., *Oxford Studies in Philosophy of Religion*, Oxford University Press.

BonJour, Laurence. (1985) *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press,

Choo, Frederick. (2018) 'The epistemic significance of religious disagreements: Cases of unconfirmed superiority disagreements' *Topoi* **40**: 1139–1147. https://doi.org/10.1007/s11245-018-9599-4

Elga, Adam. (2007) 'Reflection and disagreement' *Nous* **41**: 478–502. https://doi.org/10.1111/j.1468-0068.2007.00656.x

Elgin, Catherine. (2010) 'Persistent Disagreement', in Robert Feldman & Ted A Warfield, eds., *Disagreement*, Oxford University Press.

Enoch, David. (2010) 'Not just a truthometer: Taking oneself seriously (but not too seriously) in cases of peer disagreement' *Mind* **119**: 953–997. https://doi.org/10.1093/mind/fzq070

Feldman, Richard. (2006) 'Epistemological Puzzles about Disagreement', in Stephen C Hetherington, ed., *Epistemology futures*, Oxford University Press.

Feldman, Richard. (2007) 'Reasonable Religious Disagreements', in Louise M. Antony, ed., *Philosophers without gods: Meditations on atheism and the secular life*, Oxford University Press.

Hale, Bob, and Crispin Wright. (2001) *The reason's proper study: Essays towards a neo-Fregean Philosophy of Mathematics*. Oxford: Clarendon.

Kelly, Thomas. (2010) 'Peer Disagreement and Higher Order Evidence', in Richard Feldman & Ted A Warfield, eds., *Disagreement*, Oxford University Press.

Kelly, Thomas. (2005) 'The Epistemic Significance of Disagreement', in Tamar Z Gendler & John Hawthorne, eds., *Oxford studies in epistemology*, Clarendon Press.

King, Nathan. (2011) 'Disagreement: What's the problem? or a good peer is hard to find' *Philosophy and Phenomenological Research* **85**: 249-272 https://doi.org/10.1111/j.1933-1592.2010.00441.x

Lackey, A Jennifer. (2014) 'Taking Religious Disagreement Seriously', in Laura Frances Callahan & Timothy O'Connor, eds., *Religious faith and intellectual virtue*, Oxford University Press.

Lemos, Noah M. (2021) *An introduction to the theory of knowledge*. Cambridge University Press,

Lougheed, Kirk. (2020) *The epistemic benefits of disagreement*. Springer.

Matheson, Jonathan. (2015) *The epistemic significance of disagreement*. Palgrave Macmillan.

Oppy, Graham. (2010) 'Disagreement' *International Journal for Philosophy of Religion* **68**: 183–199. https://doi.org/10.1007/s11153-010-9254-5

Plantinga, Alvin. (2000) *Warranted Christian belief*. Oxford University Press.

Plato. (2002) *Five dialogues Euthyphro, Apology, Crito, Meno, Phaedo* in Georges Maximilien Antoine Grube & John M Cooper, Trans., Hackett.

Robustness Analysis as a Procedure for Determining Difference-Makers

Jan Zebrowski[295]

University of Cambridge

Abstract

Model-based science—the style of theoretical work dominant in many social sciences, including economics—studies complex real-world systems indirectly through highly idealized model systems. The viability of model-based science pivots on the possibility of determining 'difference-makers' for every system it studies. However, economic systems are neither clearly circumscribed nor 'closed' in the sense that any outcome studied by economists is, to a greater or lesser extent, causally influenced by an infinitely complex network of factors. This makes salient the question: How do economists determine which factors are explanatorily relevant to any given outcome and should be included in its explanation? This is the problem of explanatory relevance. In this paper, I try to make headway towards solving this problem using robustness analysis (RA)—a well-known procedure in theoretical economics by which modellers gauge the sensitivity of their models' results to assumptions that fuel the derivation of these results.

[295] Jan Zebrowski was awarded a BSc (with first class honours) in Philosophy, Logic and Scientific Method by the London School of Economics and Political Science (LSE) and is currently reading for the MPhil in History and Philosophy of Science and Medicine at the University of Cambridge. His research interests fall into three research areas: general philosophy of science, especially epistemology of science, philosophy of biology and philosophy of economics.

1. Introduction

Model-based science—the style of theoretical work dominant in many social sciences, including economics—studies complex real-world systems indirectly through highly idealized model systems.[296] Much of model-based science is about mapping the components, activities, properties, and organizational features of these model systems onto the corresponding components, activities, properties, and organizational features of their target, usually real-world, systems. If the former resemble the latter in relevant respects, these model systems are said to *represent* their target systems and may be used to explain and understand them.[297] Model-based explanations are successful, on the causal-mechanical model of explanation, if they accurately represent all factors that make a difference to their target systems and abstract away all factors that do not. The viability of model-based science, then, pivots on the possibility of determining 'difference-makers' for every system it studies.

However, economic systems are neither clearly circumscribed nor 'closed,' in the sense that any outcome studied by economists is, to a greater or lesser extent, causally influenced by an infinitely complex network of factors.[298] This makes salient the question: How do economists determine which factors are explanatorily relevant to any given outcome and should be included in its explanation? This is the problem of explanatory relevance.

[296] Godfrey-Smith, Peter (2006) "The Strategy of Model-based Science", *Biology and Philosophy*, **21**(5): 725–740; Weisberg, Michael (2006a) "Forty Years of 'The Strategy': Levins on Model Building and Idealization", *Biology and Philosophy*, **21**(5): 623–645; Weisberg, Michael (2007a) "Who Is a Modeler?", *The British Journal for the Philosophy of Science*, **58**(2): 207–233.

[297] Giere, Ronald (1988) *Explaining Science: A Cognitive Approach*. Chicago, IL: Chicago University Press; Godfrey-Smith, Peter (2009) "Models and Fictions in Science", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, **143**(1): 101–116.

[298] Davidson, Donald (1970) "Mental Events." In: L. Foster and J. W. Swanson (eds), *Experience and Theory*. Oxford: Clarendon Press, 207–224.

In this paper, I try to make headway towards solving this problem using robustness analysis (RA)—a well-known procedure in theoretical economics by which modellers gauge the sensitivity of their model's results to assumptions that fuel the derivation of these results.[299]

The paper is divided into two sections. Section 1 is expository: I explain why the problem of explanatory relevance besets the causal-mechanical model of explanation and why it persists in the face of the pragmatic dimension of explanation, such as the context in which an explanation occurs and interests of those providing and receiving it. Section 2 is polemical: I introduce the eliminative procedure for determining difference-makers and bring up two problems that, although not damning, show that it does not give us much of a handle on the problem of explanatory relevance. Then, I make a case for RA as a procedure for determining difference-makers, compare RA with the eliminative procedure and conclude that it is better geared to the style of theoretical work dominant in economics.

2. Explanation and Explanatory Relevance

2.1 The Causal-Mechanical Account of Explanation

The causal-mechanical model of explanation is based on the intuitive idea that factors cited in the explanation must fit into a causal nexus with the outcome to be explained (the explanandum-outcome).[300] It measures the success of an explanation (i) by how well it represents causal mechanisms that bring about its explanandum-outcome and (ii) by how well it discriminates the factors that make a difference to its explanandum-outcome from the rest of the factors that played a role

[299] Woodward, James (2006) "Some Varieties of Robustness", *Journal of Economic Methodology*, **13**(2): 219–240.

[300] Lewis, David (1986) "Causal Explanation" in: *Philosophical Papers*, Vol. II. Oxford: Oxford University Press; Railton, Peter (1978) "A Deductive-Nomological Model of Probabilistic Explanation", *Philosophy of Science*, **45**(2): 206–226; Salmon, Wesley C. (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

in producing it.[301] At least some explanations that score a success in this way are model-based, in the sense that they rely on highly idealized models to accurately represent the causal mechanisms that bring about their explanandum-outcomes. Such models explain if their components, activities, properties and organizational features correspond to, or map onto, the components, activities, properties and organizational features of their target systems.[302] They provide understanding of real-world phenomena not by showing "*that* [these real-world phenomena] fit into a *nomic* nexus" but by showing "*how* [these real-world phenomena] fit into a *causal* nexus."[303] But what is it to show how real-world phenomena fit into a causal nexus?

2.1.1 Woodward's Manipulationist Framework.

Woodward[304] introduces a 'manipulationist' framework for analyzing causal relationships that has the rare merit of showing causation's place in the circle of interrelated concepts that includes "cause," "counterfactual dependence," "explanation," and "explanatory relevance." The basic idea is that to show how X and Y fit into a causal nexus is to show how they fit into a pattern of counterfactual dependance. Is manipulation of X a way of manipulating Y? If changes in X produced by interventions are systematically associated with changes in Y, X and Y fit into a pattern of counterfactual dependance. If this pattern of counterfactual dependance is sufficiently invariant and continues to hold under a range of interventions in X, X and Y fit into a causal nexus. A necessary and sufficient

[301] Bechtel, William and Richardson, Robert C. (2010) *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Cambridge, MA: MIT Press; Machamer, Peter, Darden, Lindley and Craver, Carl F. (2000) "Thinking About Mechanisms", *Philosophy of Science*, **67**(1): 1–25; Strevens, Michael (2009) *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

[302] Kaplan, David M. and Craver, Carl F. (2011) "The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective*", *Philosophy of Science*, **78**(4): 601–627.

[303] Salmon, Wesley C. (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

[304] Woodward, James (2003) *Making Things Happen*. New York, NY: Oxford University Press.

condition for a relationship between X and Y to be causal is therefore that it be invariant under a range of interventions.

What about explanation and explanatory relevance? It is a point made by Hitchcock[305] that our intuitive judgements concerning explanatory relevance correspond to our judgements concerning the truth of counterfactuals. Woodward[306] takes this point further and explicitly analyzes explanatory relevance in terms of counterfactual dependence. He argues that explanations must provide true 'what-if-things-had-been-different' information: information about how changes produced by interventions in the factors cited in the explanation are systematically associated with changes in its explanandum-outcome. For example, the explanation of the length of a shadow cast by a flagpole in terms of the elevation of the sun and the height of the flagpole shows how this explanandum-outcome fits into a causal nexus with the elevation of the sun and the height of the flagpole in that it does provide true 'what-if-things-had-been-different 'information: it correctly suggests that, other things being equal, had the elevation of the sun and/or the height of the flagpole been different, the length of a shadow cast by a flagpole would have been different. It bears emphasizing that true 'what-if-things-had-been-different' information is causal information in that it is, by definition, invariant under a range of interventions: if things had been different, it would continue to hold. So, Woodward's manipulationist framework collapses the distinction between providing explanation and providing information about causal mechanisms and identifies explanatory relevance with causal relevance.

2.2 The Problem of Explanatory Relevance

The snag with Woodward's concept of explanatory relevance is that it is too weak. It is a familiar point that social systems in general and economic systems in particular

[305] Hitchcock, Christopher R. (1995) "Salmon on Explanatory Relevance", *Philosophy of Science*, **62**(2): 304–320.

[306] Woodward, James (2003) *Making Things Happen*. New York, NY: Oxford University Press.

are neither clearly circumscribed nor 'closed,' in the sense that any outcome studied by economists is, to a greater or lesser extent, causally influenced by an infinitely complex network of factors.[307] These include environmental stresses acting on individuals, gravitational forces exerted by distant stars acting on households and firms etc. And yet, when explaining economic outcomes such as, say, economic growth in the UK in the fourth quarter of 2022 (E), economists neglect environmental stresses (C_1) and gravitational forces exerted by distant stars (C_2), instead attuning their models to changes in technology (A), physical capital stock (K) and human capital stock (L). How do economists account for this neglect? Surely, changes in C_1 and C_2 produced by interventions would be systematically associated with changes in E. For example, it is a safe bet that changes in C_1 produced by interventions that increase viral mutation rates would be systematically associated with slowdowns in E. It follows that C_1 and E fit into a pattern of counterfactual dependance. If this pattern of counterfactual dependance was sufficiently invariant and continued to hold under a range of interventions in C_1, C_1 and E would fit into a causal nexus. Supposing that it was, would it be reason enough to conclude that C_1 is causally *and* explanatorily relevant to E? But then, an analogous argument could be run with C_2, C_3, …, C_n, leading to the untoward conclusion that economic models targeting E would have to be infinitely complex to accurately represent causal mechanisms underlying it. So, there are grounds for diffidence about identifying explanatory relevance with causal relevance.

2.2.1 The Pragmatics of Explanation.

It might be objected that explanation is not a two-term relation between an explanandum-outcome and a model that explains it, but a three-term relation between an explanandum-outcome, a model that explains it and a range of pragmatic factors, such as the context in which it occurs and interests of those

[307] Davidson, Donald (1970) "Mental Events." In: L. Foster and J. W. Swanson (eds), *Experience and Theory*. Oxford: Clarendon Press, 207–224.

providing and receiving it. For van Fraassen[308] and Lewis,[309] explanations are answers to why-questions and therefore require irreducible reference to such pragmatic factors. Why-questions are often explicitly contrastive: a request for the explanation of P often takes the form "Why P rather than Q?" But even when why-questions are not explicitly contrastive, they may be construed as "Why P rather than other members of its contrast class X?", where the contrast class X is a class of alternatives to P specified by the context in which they occur. Garfinkel[310] and Lipton[311] argue that what counts as an explanatorily relevant factor depends not only on P but also on its contrast class X. So, returning to our example, whether or not C_1 and C_2 are explanatorily relevant to E depends on the context in which the explanation of E occurs.

Explanation may very well be context- and interest-relative but specifying the context in which it occurs and interests of those providing and receiving it will not give us much of a handle on the problem of explanatory relevance. In most contexts, when explaining E, the influence of such spurious causal factors as C_1 and C_2 on E will be straightforwardly ruled out courtesy of pragmatic considerations. Granted. However, pragmatic considerations alone will not do much to discriminate the factors that make a difference to E from the rest of the factors that played a role in producing it. They will rule out C_1 and C_2 as explanatorily irrelevant, but they will neither rule out nor help us gauge the relative magnitude of a motley bunch of other factors that, to a greater or lesser extent, causally influence E. These other factors include "knowledge externalities" featured in Romer's[312] growth model, "learning

[308] Van Fraassen, Bas (1980) *The Scientific Image*. Oxford: Clarendon Press.

[309] Lewis, David (1986) "Causal Explanation" in: *Philosophical Papers*, Vol. II. Oxford: Oxford University Press.

[310] Garfinkel, Alan (1981) *Forms of Explanation: Rethinking the Questions in Social Theory*. New Haven, CT: Yale University Press.

[311] Lipton, Peter (1990) "Contrastive Explanation", in D. Knowles (ed.), *Explanation and Its Limits*. Cambridge: Cambridge University Press, 247–266; Lipton, Peter (1991). *Inference to the Best Explanation*. London and New York: Routledge.

[312] Romer, Paul M. (1986) "Increasing Returns and Long-Run Growth", *Journal of Political Economy*, **94**(5): 1002–1037.

externalities" featured in Tamura's[313] growth model, "international integration" featured in Rivera-Batiz and Romer's[314] growth model etc. All these factors are causally relevant to E, in the sense that, if we could but manipulate them, it is a safe bet that changes in them produced by interventions would be systematically associated with changes in E. And, in most contexts, all these factors will be explanatorily relevant to E. My worry is that since a successful explanation must accurately represent causal mechanisms that bring about its explanandum-outcome, in most contexts explaining E will involve citing a network of causal factors that, though not infinite, will be very complex and intractable indeed. And since it is unlikely that any one model will represent all of them, explaining E will likely involve a battery of models, each requiring its own set of assumptions to fuel the derivation of E. I submit that our understanding of economic outcomes would be aided greatly if we could determine precisely which factors are necessary for their causal production.

3. In Search of Difference-makers

Strevens[315] goes a long way towards addressing the problem of explanatory relevance. Like Hitchcock[316] and Woodward,[317] he argues that difference-making is a necessary condition for explanatory relevance: only factors that make a difference to an explanandum-outcome are explanatorily relevant to this outcome. However, unlike Hitchcock[318] and Woodward,[319] he offers a procedure for determining precisely which factors are necessary for the causal production of particular

[313] Tamura, Robert (1991) "Income Convergence in an Endogeneous Growth Model", *Journal of Political Economy*, **99**(3): 522–540.

[314] Rivera-Batiz, Luis A., and Romer, Paul M. (1991) "Economic Integration and Endogenous Growth", *The Quarterly Journal of Economics*, **106**(2): 531–555.

[315] Strevens, Michael (2004) "The Causal and Unification Approaches to Explanation Unified: Causally", *Noûs*, **38**(1): 154–176.

[316] Hitchcock, Christopher R. (1995) "Salmon on Explanatory Relevance", *Philosophy of Science*, **62**(2): 304–320.

[317] Woodward, James (2003) *Making Things Happen*. New York, NY: Oxford University Press.

[318] Hitchcock, Christopher R. (1995) "Salmon on Explanatory Relevance", *Philosophy of Science*, **62**(2): 304–320.

[319] Woodward, James (2003) *Making Things Happen*. New York, NY: Oxford University Press.

outcomes. He calls this procedure "the eliminative procedure for determining difference-makers."[320]

3.1 The Eliminative Procedure for Determining Difference-makers

According to Strevens,[321] difference-makers for E may be determined courtesy of the following procedure:

1. Take the infinitely complex network of factors causally relevant to E, and "find a part [of the network] that [is] in itself sufficient to causally produce E."[322]
2. Remove from it all factors that are not necessary to causally produce E. This means removing all factors that do not "play a role in the entailment of E."[323]

The eliminative procedure is disarmingly simple. A moment's consideration shows, however, that it will not go a long way towards determining difference-markers for overdetermined economic phenomena and that it runs afoul of Duhem's non-separability thesis.

Firstly, the eliminative procedure consists of two steps but I do not see exactly how the first step gets us any closer to determining difference-makers for E. Strevens[324] glosses causation in terms of the semantic relation of entailment and stipulates that a set of conditions is sufficient to causally produce E "just in case the conditions jointly entail the causal production of E." So if **C** = {C_1, C_2, …, C_n} was a set of conditions the description of which entailed the description of E, **C** would be sufficient to causally produce E. But then, I do not see exactly how causal sufficiency is different from causal relevance. My first worry is that the set of factors sufficient, in Strevens'[325] sense, to causally produce E and the set of factors causally relevant, in

[320] Strevens, Michael (2004) "The Causal and Unification Approaches to Explanation Unified: Causally", *Noûs*, **38**(1): 154–176.
[321] Ibid.
[322] Ibid.
[323] Ibid.
[324] Ibid.
[325] Strevens, Michael (2004) "The Causal and Unification Approaches to Explanation Unified: Causally", *Noûs*, **38**(1): 154–176.

Woodward's[326] sense, to E will be very similar in terms of their respective cardinalities—numbers of elements—and just as intractable.

My first worry is exacerbated by the fact that social phenomena in general and economic phenomena in particular are multiply realizable.[327] Think of all realizing conditions of economic outcomes, such as economic growth in the UK in the fourth quarter of 2022 (E). We are talking millions of economic agents making billions of transactions on a daily basis, we are talking knowledge and learning externalities, international integration etc. If that was not enough, each of these realizing conditions is itself multiply realizable. Think of all realizing conditions of these transactions, think of all realizing conditions of knowledge and learning externalities, international integration etc. Because they are multiply realizable, economic phenomena are overdetermined, in the sense that more than one set of conditions is sufficient to causally produce them. For example, E might be realized by **C** = {C_1, C_2, …, C_n}, where C_1 is the Bank of England raising interest rates to 3.50%, C_2 is the Bank of England reversing quantitive easing etc., but it might also be realized by **C′** = {C'_1, C'_2, …, C'_n}, where C'_1 is the Bank of England raising interest rates to 3.75%, C'_2 is the Bank of England not reversing quantitive easing etc. But then, if the description of both **C** and **C′** entailed the description of E, both **C** and **C′** would be sufficient, in Strevens'[328] sense, to causally produce E. And I would hazard a guess that there are more than two sets of conditions the description of which entails the description of E; such sets are a dime a dozen. If I am right, the first step of the eliminative procedure, that is, discriminating the set of factors sufficient to causally produce E from the set of factors causally relevant to E, would leave us where we started, namely with the set of factors causally relevant to E. It would not go a long way towards determining difference-markers for E.

Secondly, and more importantly, supposing that one put a finger on a set of conditions sufficient to causally produce E, it is not clear exactly how one is to

[326] Woodward, James (2003) *Making Things Happen*. New York, NY: Oxford University Press.

[327] Fodor, Jerry (1974) "Special Sciences (Or: The Disunity of Science as a Working Hypothesis)", *Synthese*, **28**(2): 97–115; Searle, John R. (1995) *The Construction of Social Reality*. London: Allen Lane; Searle, John R. (2005) "What Is an Institution?" *Journal of Institutional Economics*, **1:** 1–22.

[328] Strevens, Michael (2004) "The Causal and Unification Approaches to Explanation Unified: Causally", *Noûs*, **38**(1): 154–176.

remove from it all factors that do not "play a role in the entailment of E." It is a familiar point, known as Duhem's non-separability thesis, that the empirical content of a hypothesis or a theory or any other empirically significant unit cannot be 'parcelled out' among its parts.[329] This is because a conjunction of two hypotheses or theories or any other empirically significant units may very well entail a result that is not entailed by either of them taken in isolation.[330] For example, a sufficiently detailed list of facts about the behaviour of nerve cells in my brain entail propositions about my consciousness, i.e., about me being conscious of my body, my self, the world at large etc. However, no single fact about the behaviour of nerve cells in my brain entails such propositions. So if $\mathbf{C} = \{C_1, C_2, \ldots, C_n\}$ entails E, there is, and can be, no guarantee that C_1 alone entails E or that C_2 alone entails E or that C_n alone entails E. Strevens might reply that all factors that do not "play a role in the entailment of E" might be removed 'by hand.' If $\mathbf{C} = \{C_1, C_2, \ldots, C_n\} \vDash$ E, we might remove C_1 from **C** 'by hand' and see if the resulting set entailed E. If it did, we might conclude that C_1 is not necessary to causally produce E. The same might be done with C_2, C_3, all the way to C_n.

But then, suppose that after we had removed C_1 from **C** 'by hand,' we found that the resulting set did not entail E. Does this finding license the conclusion that C_1 is necessary to causally produce E? The answer is an emphatic "no." For all we know, C_1 might be an 'enabler' of C_2, in the sense that it might be a factor, one of many, without which C_2 is causally inert. If C_2 had other enablers, C_1 would not be necessary to causally produce E. In general, the members of **C** might 'interlock' to such an extent that removing all those that do not "play a role in the entailment of E"

[329] See Ariew, Roger (1984) "The Duhem Thesis", *The British Journal for the Philosophy of Science*, **35**(4): 313–325; Duhem, Pierre (1917) "Liste des Publications de P. Duhem" and "Notice sur les Travaux Scientifiques de Duhem," *Mémoires de la Société des Sciences Physiques et naturelles de Bordeaux*, 7, 41–169. English translation of Parts 2 and 3 of "Notice" in Duhem (1996); Duhem, Pierre (1954) *The Aim and Structure of Physical Theory*. Princeton, NJ: Princeton University Press; Quinn Philip L. (1974) "What Duhem Really Meant", in: *Methodological and Historical Essays in the Natural and Social Sciences*. Dordrecht: Springer.

[330] Quine, Willard V. O. (1981) *Theories and Things*. Cambridge, MA: Harvard University Press.

might be little more than guesswork.[331] This is why I am skeptical about the reply that all factors that do not "play a role in the entailment of E" might be removed 'by hand.'

I argued that the eliminative procedure will not go a long way towards determining difference-markers for overdetermined economic phenomena and that it runs afoul of Duhem's non-separability thesis. I grant that these problems are not damning. For one thing, not all phenomena are as overdetermined as economic phenomena. For another, Duhem's non-separability thesis does not always hold. It was floated with theoretical physics in mind and applies only to those sciences that do not "observe facts directly, but substitute for them measurements [...] of magnitudes that only a mathematical theory has defined."[332] So the eliminative procedure may apply to some cases. My point is that, although it is advertised as "a procedure for determining difference-makers," it does not apply 'across the board' to all cases and therefore does not give us much of a handle on the problem of explanatory relevance.

3.2 Robustness Analysis as a Procedure for Determining Difference-makers

In this Subsection, I try to make headway towards solving the problem of explanatory relevance using robustness analysis (RA)—a well-known procedure in theoretical economics by which modellers gauge the sensitivity of their models' results to assumptions that fuel the derivation of these results. I compare RA with the eliminative procedure and conclude that it is better geared to the style of theoretical work dominant in economics. Then, in Subsections 2.2.1 and 2.2.2, I consider two objections to RA as a procedure for determining difference-makers and try to meet them.

[331] Quine, Willard V. O. (1981) *Theories and Things*. Cambridge, MA: Harvard University Press.

[332] Duhem, Pierre (1917) "Liste des Publications de P. Duhem" and "Notice sur les Travaux Scientifiques de Duhem," *Mémoires de la Société des Sciences Physiques et naturelles de Bordeaux*, 7, 41–169. English translation of Parts 2 and 3 of "Notice" in Duhem (1996).

Recall that economics is a model-based science in that it studies complex real-world systems indirectly through highly idealized model systems. These model systems are "highly idealized" in that descriptions specifying them contain (descriptively false) idealizing assumptions introduced to isolate causal mechanisms underlying their target systems and (descriptively false) tractability assumptions introduced for reasons of mathematical tractability.[333] For example, the description specifying Howitt's[334] growth model contains the assumption that labour force growth rates are equal across countries. This assumption is descriptively false: labour force growth rates are not equal across countries. For example, according to CEIC data, Bangladesh's labour force participation rate increased to 58.8% in Dec 2022, compared with 58.2% in the previous year; by contrast, Bahrain's labour force participation rate increased to 71.7% in Dec 2022, compared with 71.0% in the previous year. But then, the assumption that labour force growth rates are equal across countries is introduced "not because [it is thought] accurate for describing what is happening now, but because [it is thought] a convenient fiction for a steady state model to explore international spillovers."[335]

Depending on how they are specified, different models maximize different theoretical desiderata of model building. Weisberg[336] argues that there are three such desiderata, namely generality, realism and precision, and shows that they trade off against each other. For example, some models are more general than others, in the sense that they can be applied to more target systems than others. However, generality trades off against precision: general models are those that leave vague the

[333] Wimsatt, William C. (1987) "False Models as a Means to Truer Theories", in: M. Nitecki and A. Hoffmann (eds), *Neutral Models in Biology*. Oxford: Oxford University Press, 23–55.

[334] Howitt, Peter (2000) "Endogenous Growth and Cross-Country Income Differences", *The American Economic Review*, **90**(4): 829–846.

[335] Klenow, Peter J. and Rodriguez-Clare, Andrés (2005) "Externalities and Growth", In: P. Aghion and S. Durlauf (eds), *Handbook of Economic Growth*. North-Holland: Elsevier, 817–861.

[336] Weisberg, Michael (2003) "When Less is More: Tradeoffs and Idealization in Model Building." Dissertation, Stanford University; Weisberg, Michael (2006a) "Forty Years of 'The Strategy': Levins on Model Building and Idealization", *Biology and Philosophy*, **21**(5): 623–645.

magnitude of the causal forces they describe. The three-way trade-off between generality, realism and precision explains why economic models are in no short supply and why model-based explanations enjoy such a wide currency in economics. And I submit that it is this feature of economic models—abundance—that makes it particularly worthwhile to analyze their results for robustness.

What does it mean to analyze the results of economic models for robustness? Let **M** = {M_1, M_2, …, M_n} be a set of diverse models of some economic phenomenon, say, economic growth. Suppose that each member of **M** can be broken down into a causal core (C_n) and a belt of auxiliary assumptions (A_1, A_2, …, A_n). And suppose that two members of **M** are diverse if, and only if, they can be broken down into logically non-equivalent auxiliary assumptions. Given that **M** is a set of *diverse* models, we know that each of its members contains logically non-equivalent assumptions and maximizes different theoretical desiderata of model building so our background knowledge does not favour any member of **M** over its competitors. In a classic paper, Levins[337] argues that if these models, despite their logically non-equivalent assumptions, converge on similar results, this would justifiably increase the modeller's confidence that their converging on similar results depends not on the details of their assumptions but on the 'essentials' shared across them:

> [I]f these models, despite their different assumptions, lead to similar results, we have […] a robust theorem that is relatively free of the details of the model. Hence, our truth is the intersection of independent lies.[338]

A "robust theorem" is Levins'[339] term of art for a conditional statement, sometimes prefaced by a qualifying *ceteris paribus* clause, that links the 'essentials' shared across a set of diverse models with the results they converge on. So analyzing the results of economic models for robustness means searching for such theorems.[340]

[337] Levins, Richard (1966) "The Strategy of Model Building in Population Biology", *American Scientist*, **54**(4): 421–431.

[338] Ibid.

[339] Ibid.

[340] Weisberg, Michael (2006b) "Robustness Analysis", *Philosophy of Science*, **73**(5): 730–742.

Examples of robust theorems are few and far between, but consider the following one. The Classical Growth Model (M_1), the Malthusian Growth Model (M_2) and the Solow Growth Model (M_3) are members of **M**. They are diverse, in the sense that they contain logically non-equivalent assumptions, some of which are descriptively false: M_1 assumes saving-investment equality, M_2 assumes that the standard of living and the total fertility rate are directly proportional and M_3 assumes full employment of capital and labour. If M_1, M_2 and M_3, despite their logically non-equivalent assumptions, converged on similar results, say, a particular value of economic growth in the UK in the fourth quarter of 2022 (E), this convergence would be either an artefact of some causal structure shared across them or "a remarkable coincidence."[341] So after determining that M_1, M_2 and M_3 converge on a particular value of E, they might be analyzed for a common causal structure. What M_1, M_2 and M_3 have in common is that they all construe economic growth (Y) as a function of three variables: technology (A), physical capital stock (K) and human capital stock (L). So it looks like we might have latched onto a robust theorem: Y is a function of A, K and H, and, *ceteris paribus*, changes in A, K and L are associated with changes in Y. Notice robust theorems are (qualified) patterns of counterfactual dependance: to say that changes in A, K and L are associated with changes in Y is to say that A, K and L fit into a pattern of counterfactual dependance with Y.

I therefore submit that RA lends itself admirably to determining the factors that make a difference to a given explanandum-outcome and distinguishing these difference-makers from the rest of the factors that played a role in producing it. Although it would be wishful thinking to assume that RA is a sure-fire procedure for determining difference-makers, I think that the odds are good that the 'essentials' shared across a set of diverse models of some economic phenomenon (E) are factors that matter, counterfactually, to E: if we could but manipulate them, this would be a way of manipulating E. And I think that adding new members to this set would increase these odds: the greater the cardinality—number of elements—of a set of diverse models, the greater the odds that the 'essentials' shared across them fit into a

[341] Kuorikoski, Jaakko. et al. (2010) "Economic Modelling as Robustness Analysis", *The British Journal for the Philosophy of Science*, **61**(3): 541–567.

pattern of counterfactual dependance with E. How does RA compare with the eliminative procedure? I argued that the eliminative procedure will not go a long way towards determining difference-markers for economic phenomena because they are overdetermined, in the sense that more than one set of conditions is sufficient to causally produce them. What I argued is a crippling handicap for the eliminative procedure is a valuable asset to RA. Overdetermination of economic phenomena ensures the steady supply of diverse models targeting them. If these models, despite their logically non-equivalent assumptions, converged on similar results, they might be analyzed for a common causal structure.

3.2.1 Objection 1: Models Do Not Decompose That Way.

I also argued that the eliminative procedure runs afoul of Duhem's non-separability thesis. One might balk at RA and object that it does so too. I acknowledge this objection and think it is well-taken. There is no denying that analyzing diverse models for a common causal structure assumes that models can be broken down into parts.[342] However, I think there is a difference between RA and the eliminative procedure. Showing that a set of diverse models converges on a particular result (E), analyzing these models for a common causal structure (C) and then attributing E to C is a way of 'parcelling out' the empirical content of a model among its parts. Granted. However, I argued that it would be wishful thinking to assume that RA is a sure-fire procedure for determining difference-makers. It is not. C may be a difference-maker for E or not, but it is possible that it is. My contention, then, is that RA allows us to make educated guesses about difference-makers in the face of Duhem's non-separability thesis. By contrast, Strevens[343] argues that "in order for an event C to qualify as a difference-maker for an event E, it is necessary and sufficient that C appear in a kernel Strevens'[344] term of art for the end product of the eliminative procedure] for E." Appearing in a "kernel" for E may be sufficient for

[342] Rice, Collin (2019) "Models Don't Decompose That Way: A Holistic View of Idealized Models", *British Journal for the Philosophy of Science*, **70**(1): 179–208.

[343] Strevens, Michael (2004) "The Causal and Unification Approaches to Explanation Unified: Causally", *Noûs*, **38**(1): 154–176.

[344] Ibid.

qualifying as a difference-maker for E, but it is not necessary. Duhem's non-separability thesis, as long as it holds, gives the lie to this claim. One cannot strip a model of all parts that do not "play a role in the entailment of E." This is because these parts might 'interlock' to such an extent that, removing any one of them would make the model causally inert and its empirical consequence class empty. So, if appearing in a "kernel" for E was necessary for qualifying as a difference-maker for E, I have a hunch that no event would qualify as a difference-maker for E.

3.2.2 Objection 2: Models Do Not Represent Causal Mechanisms.

One might also object that since economic models are shot through with false assumptions, they misrepresent their target systems and therefore are not explanatory. For example, Odenbaugh and Alexandrova[345] argue that if a model contains false assumptions, the causal mechanism it represents "cannot be what the model says it is." Similarly, Alexandrova and Northcott[346] argue that models "do not qualify as causal explanations because they are false and therefore do not identify any actual causes." But then, it is a familiar point that *all* models contain false assumptions and are false in this sense. If only true models could identify actual causes, this would lead to the untoward conclusion that none of them identifies actual causes and therefore none of them qualifies as a causal explanation. Indeed, Kuorikoski et al.[347] argue by *reductio ad absurdum* (RED) that if every false assumption entering a given model-based explanation had to be discharged or de-idealized with some true assumption for it to accurately represent causal mechanisms, no model-based explanation, but only "reality itself," would be "capable of such a feat." And yet, the argument goes, at least some model-based explanations are explanatory.

[345] Odenbaugh, Jay and Alexandrova, Anna (2011) "Buyer Beware: Robustness Analyses in Economics and Biology", *Biology and Philosophy*, **26**(5): 757–771.

[346] Alexandrova, Anna and Northcott, Robert (2013) "It's Just A Feeling: Why Economic Models Do Not Explain", *Journal of Economic Methodology*, 2**0**(3): 262–267.

[347] Kuorikoski, Jaakko. et al. (2012) "Robustness Analysis Disclaimer: Please Read the Manual Before Use!", *Biology and Philosophy*, **27**(6): 891–902.

Kuorikoski et al.[348] conclude, by RED, that it is not the case that every false assumption entering a given model-based explanation has to be discharged or de-idealized with some true assumption for it to accurately represent causal mechanisms.

I grant that if no model represented causal mechanisms, RA as a procedure for determining difference-makers would be a 'no go.' But the claim that models containing falsehoods do not represent causal mechanisms strikes me as plain wrong. Hausman[349] and Mäki[350] argue that *all* models misrepresent their targets, but this does not automatically bar them from representing causal mechanisms and being explanatory. More specifically, the fact that a model contains falsehoods "does not preclude employing the model in giving explanations if the explanations do not rely on the falsehoods."[351] Since it is not always the case that explanations in economics rely on the falsehoods contained in models, the premise that only true models can identify actual causes and be explanatory is not a good one. The explaining may very well be done by accurate claims about causal mechanisms, should models contain any. In other words, models containing falsehoods may accurately represent *some* causal mechanisms underlying their targets, and therefore they may be used to explain their targets, insofar as these false assumptions are not "driving the results."[352]

4. Conclusion

RA is a well-known procedure in theoretical economics by which modellers gauge the sensitivity of their models' results to assumptions that fuel the derivation of these results. I argued that the applicability of RA is wider than that and made a case

[348] Ibid.

[349] Hausman, Daniel M. (2013) "Paradox Postponed", *Journal of Economic Methodology*, **20**(3): 250–254.

[350] Mäki, Uskali (2013) "On a Paradox of Truth, or How Not to Obscure the Issue of Whether Explanatory Models Can Be True", *Journal of Economic Methodology*, **20**(3): 268–279.

[351] Hausman, Daniel M. (2013) "Paradox Postponed", *Journal of Economic Methodology*, **20**(3): 250–254.

[352] Kuorikoski, Jaakko. et al. (2010) "Economic Modelling as Robustness Analysis", *The British Journal for the Philosophy of Science*, **61**(3): 541–567.

for RA as a procedure for determining difference-makers for overdetermined economic phenomena. Although RA is not a sure-fire procedure for determining difference-makers, it allows us to make educated guesses about what factors matter, counterfactually, to a given explanandum-outcome. It also compares favorably with the eliminative procedure, though, to be sure, both procedures are beset with similar problems, especially Duhem's non-separability thesis, and neither of them applies 'across the board' to all cases. In general, I think RA may have the edge over the eliminative procedure when the explanandum-outcome is a *type*-phenomenon. It may give us a handle on the problem of explanatory relevance when we want to determine difference-makers for economic growth *in general* or inflation *in general*. By contrast, the eliminative procedure may have the upper hand when the explanandum-outcome is a *token*-phenomenon. It may give us a handle on the problem of explanatory relevance when we want to determine difference-makers for a *particular* economic growth rate or a *particular* inflation rate.

References

Alexandrova, Anna and Northcott, Robert (2013) "It's Just A Feeling: Why Economic Models Do Not Explain", *Journal of Economic Methodology*, **20**(3): 262–267.

Ariew, Roger (1984) "The Duhem Thesis", *The British Journal for the Philosophy of Science*, **35**(4): 313–325.

Bechtel, William and Richardson, Robert C. (2010) *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Cambridge, MA: MIT Press.

Davidson, Donald (1970) "Mental Events." In: L. Foster and J. W. Swanson (eds), *Experience and Theory*. Oxford: Clarendon Press, 207–224.

Duhem, Pierre (1917) "Liste des Publications de P. Duhem" and "Notice sur les Travaux Scientifiques de Duhem," *Mémoires de la Société des Sciences Physiques et naturelles de Bordeaux*, 7, 41–169. English translation of Parts 2 and 3 of "Notice" in Duhem (1996).

Duhem, Pierre (1954) *The Aim and Structure of Physical Theory*. Princeton, NJ: Princeton University Press.

Duhem, Pierre (1996) *Essays in History and Philosophy of Science*. Indianapolis, IN: Hackett.

Garfinkel, Alan (1981) *Forms of Explanation: Rethinking the Questions in Social Theory*. New Haven, CT: Yale University Press.

Giere, Ronald (1988) *Explaining Science: A Cognitive Approach*. Chicago, IL: Chicago University Press.

Godfrey-Smith, Peter (2006) "The Strategy of Model-based Science", *Biology and Philosophy*, **21**(5): 725–740.

Godfrey-Smith, Peter (2009) "Models and Fictions in Science", *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, **143**(1): 101–116.

Fodor, Jerry (1974) "Special Sciences (Or: The Disunity of Science as a Working Hypothesis)", *Synthese*, **28**(2): 97–115.

Hausman, Daniel M. (2013) "Paradox Postponed", *Journal of Economic Methodology*, **20**(3): 250–254.

Hesslow, Germund (1983) "Explaining Differences and Weighting Causes", *Theoria*, **49**(2): 87–111.

Hesslow, Germund (1988) "The Problem of Causal Selection", In: D. J. Hilton (ed.), *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*. New York, NY: New York University Press.

Hitchcock, Christopher R. (1995) "Salmon on Explanatory Relevance", *Philosophy of Science*, **62**(2): 304–320.

Howitt, Peter (2000) "Endogenous Growth and Cross-Country Income Differences", *The American Economic Review*, **90**(4): 829–846.

Kaplan, David M. and Craver, Carl F. (2011) "The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective*", *Philosophy of Science*, **78**(4): 601–627.

Klenow, Peter J. and Rodriguez-Clare, Andrés (2005) "Externalities and Growth", In: P. Aghion and S. Durlauf (eds), *Handbook of Economic Growth*. North-Holland: Elsevier, 817–861.

Kuorikoski, Jaakko. et al. (2010) "Economic Modelling as Robustness Analysis", *The British Journal for the Philosophy of Science*, **61**(3): 541–567.

Kuorikoski, Jaakko. et al. (2012) "Robustness Analysis Disclaimer: Please Read the Manual Before Use!", *Biology and Philosophy*, **27**(6): 891–902.

Levins, Richard (1966) "The Strategy of Model Building in Population Biology", *American Scientist*, **54**(4): 421–431.

Lewis, David (1986) "Causal Explanation" in: *Philosophical Papers*, Vol. II. Oxford: Oxford University Press.

Lipton, Peter (1990) "Contrastive Explanation" in: D. Knowles (ed.), *Explanation and Its Limits*. Cambridge: Cambridge University Press, 247–266.

Lipton, Peter (1991) *Inference to the Best Explanation*. London and New York: Routledge.

Machamer, Peter, Darden, Lindley and Craver, Carl F. (2000) "Thinking About Mechanisms", *Philosophy of Science*, **67**(1): 1–25.

Mäki, Uskali (2013) "On a Paradox of Truth, or How Not to Obscure the Issue of Whether Explanatory Models Can Be True", *Journal of Economic Methodology*, **20**(3): 268–279.

Odenbaugh, Jay and Alexandrova, Anna (2011) "Buyer Beware: Robustness Analyses in Economics and Biology", *Biology and Philosophy*, **26**(5): 757–771.

Quine, Willard V. O. (1981) *Theories and Things*. Cambridge, MA: Harvard University Press.

Quinn Philip L. (1974) "What Duhem Really Meant", in: *Methodological and Historical Essays in the Natural and Social Sciences*. Dordrecht: Springer.

Railton, Peter (1978) "A Deductive-Nomological Model of Probabilistic Explanation", *Philosophy of Science*, **45**(2): 206–226.

Rice, Collin (2019) "Models Don't Decompose That Way: A Holistic View of Idealized Models", *British Journal for the Philosophy of Science*, **70**(1): 179–208.

Rivera-Batiz, Luis A., and Romer, Paul M. (1991) "Economic Integration and Endogenous Growth", *The Quarterly Journal of Economics*, **106**(2): 531–555.

Romer, Paul M. (1986) "Increasing Returns and Long-Run Growth", *Journal of Political Economy*, **94**(5): 1002–1037.

Salmon, Wesley C. (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Searle, John R. (1995) *The Construction of Social Reality*. London: Allen Lane.

Searle, John R. (2005) "What Is an Institution?" *Journal of Institutional Economics*, **1:** 1–22.

Strevens, Michael (2004) “The Causal and Unification Approaches to Explanation Unified: Causally”, *Noûs*, **38**(1), 154–176.

Strevens, Michael (2009) *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Tamura, Robert (1991) “Income Convergence in an Endogeneous Growth Model”, *Journal of Political Economy*, **99**(3): 522–540.

Van Fraassen, Bas (1980) *The Scientific Image*. Oxford: Clarendon Press.

Weisberg, Michael (2003) “When Less is More: Tradeoffs and Idealization in Model Building”. Dissertation, Stanford University.

Weisberg, Michael (2006a) “Forty Years of ‘The Strategy’: Levins on Model Building and Idealization”, *Biology and Philosophy*, **21**(5): 623–645.

Weisberg, Michael (2006b) “Robustness Analysis”, *Philosophy of Science*, **73**(5): 730–742.

Weisberg, Michael (2007a) “Who Is a Modeler?”, *The British Journal for the Philosophy of Science*, **58**(2): 207–233.

Weisberg, Michael (2007b) “Three Kinds of Idealization”, *The Journal of Philosophy*, **104**(12): 639–659.

Wimsatt, William C. (1987) “False Models as a Means to Truer Theories”, in: M. Nitecki and A. Hoffmann (eds), *Neutral Models in Biology*. Oxford: Oxford University Press, 23–55.

Woodward, James (2003) *Making Things Happen*. New York, NY: Oxford University Press.

Woodward, James (2006) “Some Varieties of Robustness”, *Journal of Economic Methodology*, **13**(2): 219–240.

Founded in 2019, the *Undergraduate Philosophy Journal of Australasia* (UPJA) is the first undergraduate philosophy journal run by students from Australasia. We publish one volume and host two conferences annually and interview philosophers with a substantial connection to Australasia. We aim to be an inclusive and diverse journal and welcome submissions from undergraduates (and recent graduates) worldwide, on any philosophical topic, so long as the author attempts to make a substantive contribution to contemporary philosophy. Submissions from women and other members of underrepresented groups in philosophy, including those for whom English is not their first language, are particularly encouraged.