On the Use of Thought Experiments in the Metaphysics of Race

by

Armin Mirsanaye

Submitted in partial fulfilment of the requirements
for the degree of Master of Arts

at

Dalhousie University
Halifax, Nova Scotia
April 2024

In friendship with the titan of wisdom

Table of Contents

List of Tables

Abstract

In the metaphysics of race, philosophers continue to rely on the use of thought experiments even as they openly disagree about how they should be conducted. This is a methodological problem and, in my thesis, I argue how it could be resolved. My solution is simple: treat thought experiments as games of make-believe understood as epistemically valuable imaginings generated in response to fictional narratives. By drawing on Letitia Meynell's argument for a new account of the content of thought experiments, I adopt Kendall Walton's theory from *Mimesis as Make-Believe* to understand texts that represent fictional narratives as props prescribing imaginings to readers. Thought experiments are represented by fictional narratives, too. Walton's theory shows how this works in detail, allowing for the isolation of methodological problems in thought experiments without resorting to metaphysically loaded explanations.

List of Abbreviations and Symbols Used

OMB	The United States government's Office of Management and Budget
HPC	Homeostatic property cluster theory
H_2O	The common form of water
D_2O	Heavy water
DDW	Light water

Statement

This thesis is about the use of thought experiments in the metaphysics of race. It is *not* a thesis about the metaphysics of race. A significant portion of this thesis is indeed dedicated to topics studied in the metaphysics of race, but this is only to set the context for the thought experiments I use as case studies. The reader should consider this an advance notice.

Regarding thought experiments, the theory that I advocate for is an adaptation of Kendall Walton's theory of the representational arts. Walton's theory is *not* a theory of thought experiments. However, it is clear from his writings that thought experiments are narratives that he considers to be representational objects. Thus, as a subtype of representational objects, the focus of the theory could be brought to them. Here, I have adjusted the focal point of the theory to capture the methodology of thought experiments with added customizations. Moreover, I have made sure to distinguish my own contributions to differentiate them from Walton's concepts. The reader is advised to keep note of my modifications to prevent equivocating between my work and Walton's theory.

The argument advanced in this thesis rests on an assumption about the epistemic value of thought experimentation. This assumption could be supported by the theoretical tools developed in the thesis, but that is beyond the scope of this project. Here, it suffices to argue *how* thought experiments can be epistemically valuable or how they can fail to be epistemically valuable, rather than to argue *that* thought experiments can be epistemically valuable.

The argument in this thesis takes thought experiments to be make-believe. To make-believe, one needs to imagine. If thought experiments can be epistemically valuable, and thought experiments rely on imagination, then imagination can be epistemically valuable. If so, perhaps there could be a way to differentiate between imaginings that are epistemically valuable and imaginings that are not. Indeed, such a difference is an ordinary distinction made in various ways, such that in response to an unrealistic plan a common reply in English is “keep on dreaming!” rather than “keep on imagining!” though in practical terms both relay the same message. Somehow the association of ‘dreaming’ with ‘sleeping’ diminishes its epistemic value. But, strictly speaking, this is a difference between dreaming and imagining, rather than two ways of imagining. However, in some other languages, such as Arabic and Persian, or the languages influenced by the two, similar distinctions are commonplace. False imaginations or delusions are called ‘wahm/vahm’ (وهم) and imaginations which could be epistemically valuable are called ‘khayāl/khiyāl’ (خيال). In the absence of such a linguistic convention, I continue to awkwardly phrase sentences with different permutations of ‘epistemically valuable imaginings’ in this thesis.

Regarding the metaphysics of race, I have dedicated two chapters to clarifying the debate as I understand it. Chapter Four surveys five of the most prominent contemporary positions in the metaphysics of race and catalogues them in response to a set of purely metaphysical questions. To prepare the reader for the technical nuances expressed in these questions, Chapter Three provides the working definitions and explanations of related metaphysical concepts. Neither chapter completely answers the questions raised nor offers the totality of questions which could be raised. Yet, Chapter Four is abundant with details about the metaphysics of race debate considered, while Chapter Three contains many details about the technicalities of the field of metaphysics in connection to this thesis. Even though I believe that readers well-versed in the metaphysical concepts at play should be able to skip past the two chapters with ease, I abstain from recommending it to the reader. Metaphysics and the metaphysics of race are controversial topics, both theoretically and politically. This is why I have gone to great lengths in the two chapters to clarify as much as I can, and to show the lack of clarity where I cannot.

I have tried my best to consider each position in the metaphysics of race debate neutrally and without undue prejudice. I have assumed that, in all likelihood, each position defends a reasonable metaphysical possibility to consider and each position could be supported through valid argumentation, though there may be mistakes within the current attempts at argumentation. Moreover, in fairness, my dealings with the positions considered do not presume any of them to be racist, nor as motivating or motivated by racism. At least some of the arguments considered may be racist in some conceivable way, but studying how they could be interpreted to be racist is not the aim of this thesis.

Given my fallibility, prejudices, biases and blind spots, it may be of interest to the reader if I disclose my stance toward the metaphysics of race debate that I consider in this thesis and what my view is about race or racism. About the metaphysics of race debate, as I point out in Chapter Four, I believe too much is left unaddressed by the theorists to allow me to side with one theory. So, as it stands, I can only find sensible a pluralism of compatible views, except anti-realism (broadly the view that there are no races) because of its incompatibility with such pluralism, all other positions in the debate and the contemporary common sense views in general. This being said, I have made sure not to make anti-realism into a strawman position. About race and racism, I can only say that they come in various forms and appear to be relentless realities of life in many societies today. Sadly, I suspect that theorizing about race or racism without studying the diversity of its forms is the cause of the philosophical tendency to fall short of properly capturing its lived reality.

The consequence of following certain anti-realist arguments is thinking that race or racism is the by-product of using racial language and concepts, which is to take all other positions, constructionist or not, as unwitting perpetuators of racial thinking and racism for their commitment to the language or concept. Such a view is rash enough to see this very thesis in the same light, even if no theory of race is being advocated here. So, it only stands to reason for me to think over such controversial questions, and amongst several strong contenders, that it is highly unlikely for one position to be entirely correct and others erroneous, but that it is the only ethically responsible position while all others are somehow racist. Because of this, not only do I dismiss the presumptive self-advocacy of any position, but such presumptuousness draws my attention to the diligence paid in support of any such position.

Some readers may hold that at least the realist positions that make room for a biological account must be somehow wrong or racist. They may even believe it to be a historical fact that racist doctrines always set their foundations upon biological footings. The reader should note that both of the constructionist positions in this debate make room for a biological account of race. One of them finds it plausible that a non-essentialist naturalism could be the best biological explanation, and the other explains biological differences as the by-product of biological mechanisms being historically regulated by cultural practices involved in family formation and reproduction. Subsequently, leaving aside anti-realism, every position makes room for a biological explanation in this debate.

Still, some readers may be convinced that unless a theory views its biological account as the by-product of a social or cultural explanation, it flirts with racism if racist doctrines are always founded upon biological explanations. But racist doctrines are not always founded upon biological explanations, and debating otherwise is not a philosophical position to argue for, but a fact about racist doctrines and the history of racism. So, for those who disagree with the claim, I can do no better than to suggest that they study the rivalry between the schools of ‘spiritual racism’ and ‘biological racism’ in Fascist Italy.

Acknowledgements

My deepest thanks go to my family, above all to my mother, for their continuous support. Studying philosophy is a privilege, and I owe it largely to them. Furthermore, a significant portion of the energy invested in this project is attributed to my inspiring partner, Jessie.

I am exceptionally grateful for all the help I have received from my supervisor, Dr. Duncan MacIntosh, whose professional aptitude is unmatched by any other. I thank my committee members, Dr. Greg Scherkoske and Dr. Tyler Hildebrand, for offering their time and expertise. Also, I thank Dr. Letitia Meynell for her guidance in the early stages of this project, and Dr. Mike Hymers for his influence on my studies.

I wish to extend my warmest appreciation to Dr. Brittany Kraus for her comradery and support throughout the years. I also want to express my sincerest gratitude to Dr. Colin Mitchell and Dr. Kiran Banerjee for their advice, time and felt humanity when it was in short supply.

Chapter 1: Introduction

This thesis is motivated by a seemingly simple observation: philosophers continue to rely on the use of thought experiments in the metaphysics of race as they openly disagree about how thought experiments should be conducted. So, I ask: is this a methodological problem? My answer is affirmative, and I argue for one way this methodological problem could be resolved. My solution is simple: treat thought experiments as games of make-believe understood as epistemically valuable imaginings generated in response to fictional narratives.

The power to imagine is an undeniable human asset. Tasks such as learning, problem-solving, memorizing and assessing are regularly aided by imagination. Thought experiments, counterfactuals, hypothetical examples, models, maps, blueprints, diagrams and other such epistemic devices in philosophy and science are only made possible through the power of imagination. The metaphysics of race, as the theoretical crossroad of natural and social sciences, is no different for its reliance on imagination to conduct research. However, as a subfield of social metaphysics, its scant epistemic resources often seem to be overcompensated by appeals to thought experiments.

The disagreement over the use of thought experiments in arguments about race is openly acknowledged to be a problem within the field. Indeed, any comprehensive literature review of this field of study would find that a significant portion of the discourse consists of offering, reacting to and rebutting thought experiments. Meanwhile, the lack of an accepted methodology has not dissuaded philosophers of race from continued use of the device in argumentation. Hence, the metaphysics of race is heavily invested in a method of research that it fails to methodologically account for.

In this thesis, I take it for granted that thought experiments can be epistemically useful in various ways and focus my energy on investigating why thought experimentation often fails to achieve its epistemic aims in the metaphysics of race. If there was a theory that not only explained how the narratives that shape thought experiments become epistemically useful, but was also capable of identifying how contentious thought experiments go wrong, there would be a way to both understand the epistemic function of this method and a framework for diagnosing its misapplication. Thankfully, such a theory is available.

By drawing on Letitia Meynell's argument in "Imagination and Insight: A New Account of the Content of Thought Experiments" (2014), I adopt Kendall Walton's account of representation from *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (1990) as my theoretical framework. Walton views the text representing a fictional narrative as a prop that prescribes imaginings to readers (pp. 21-24). Thought experiments, too, are represented by fictional narratives that prescribe imaginings to participants. But what sets thought experiments apart is how they are fictions with content that serves unique epistemic purposes (Meynell, 2014, p. 4150). Thus, thought experiments go wrong when their content does not achieve its planned epistemic purpose. Walton's theory shows how this works in detail.

1.1: The Plan

In the second chapter of this thesis, I present a selection of thought experiments as my case studies based on examples from Joshua Glasgow and Chike Jeffers. Numerous other examples are available in the literature, but I believe my selection conveniently exhibits the most important methodological issues to be addressed. I begin with some context about thought experiments in the metaphysics of race and then present what I have selected: I introduce six thought experiments from Glasgow and two thought experiments from Jeffers.

In the third chapter, I set the stage for my main argument by providing an overview of the nomenclature in the metaphysics of race. I begin with some general notes about the field of metaphysics, distinguishing 'Metaphysics' from 'Ontology' as it concerns my project. Then, I detail important clarifications regarding some rudimentary, but pivotal, terminology. Next, I draw attention to the metaphysical study of social reality and, further, discuss what 'Social Kinds' are meant to be.

In the fourth chapter, I focus on the central question in the metaphysics of race: *what is race*? I do not provide an answer of my own, but I outline a selection of prominent positions which have recently debated one another. This selection includes the camps of racial *Realism*, represented by Quayshawn Spencer's *Non-Essential Naturalism*, Michael Hardimon's *Minimalism* and *Populationism*, Sally Haslanger's *Socio-Political Constructionism*, and Chike Jeffers' *Cultural Constructionism*, along with the camp of racial *Anti-Realism* represented by Joshua Glasgow's *Eliminativism*. Then, I recapitulate this exposition with a critical overview.

In the fifth chapter, I detail the theoretical framework that I later apply to the thought experiments listed in Chapter One. To begin, I introduce Kendall Walton's project and delineate his theory of make-believe in six steps. First, I provide a general explanation of how the theory works and how it applies to thought experiments. Second, I discuss how Walton understands the notion of 'mimesis' by tracing it back to Ernst Gombrich. Third, I submit that Walton's theory of make-believe is an upgraded variant of

Reader-Response theory from the field of literary criticism. Fourth, I provide a working guide to ‘make-believe’ in the context of studying thought experiments. Fifth, I go over some of the key components of Walton’s theory that he refers to as the ‘Machinery of Generation,’ followed by revisions in applying them to thought experiments. Sixth, I go over the operations of Walton’s ‘Machinery of Generation,’ called the ‘Mechanics of Generation,’ followed by revisions in applying them to the inferential moves associated with thought experiments.

In the sixth chapter, I take the theoretical framework from Chapter Five and apply it to the thought experiments that I introduced in the first chapter. I take key concepts from Walton and deploy them in treatments of the case studies. Using the ‘Machinery of Generation,’ I show how specific problems arise in the thought experiments; first, with the ‘Reality Principle’; then, with the ‘Mutual Belief Principle’; and, third, with ‘authorized’ imaginings. Then, deploying the ‘Mechanics of Generation,’ I review the thought experiments suffering from faulty endogenous inferences, followed by cases suffering from neglected exogenous inferences.

The seventh chapter serves as my conclusion. I provide an overview of what I accomplish in the thesis, including questions for future study and comment on the potential limitations of the make-believe framework. I believe that my argument successfully shows that the best way to understand thought experiments is as imaginings generated by participants in response to fictional narratives, and that the methodological disagreements over their use in the metaphysics of race can be resolved by the adoption of my methodological approach in preventing misapplication.

Finally, ‘Appendix A’ offers a list of ‘Key Terms’ from a selection of concepts presented in Chapter Five. Walton offers a variety of neologisms and technical distinctions in giving his theory which could make it difficult for the reader to keep track and follow along. This glossary should help ease the reader’s burden.

Chapter 2: Thought Experiments in the Metaphysics of Race

In the metaphysics of race, there is a lack of consensus about what, if anything, makes thought experimentation acceptable or successful. Some philosophers doubt even the very legitimacy of thought experimentation (Spencer, 2019b, pp. 232-238). All the while, the literature is superabundant with thought experiments. Such a state of affairs leaves clarity about the methodology wanting. How is the content of thought experiments to be determined, and how can good inferences be made based on such content?

To illustrate, I provide eight thought experiments which I suspect readers will easily find questionable. I also include some of the ways that prominent philosophers have criticized them, as well as some of the replies to these criticisms.

2.1: Joshua Glasgow

In this section, the works of Joshua Glasgow are introduced. I review them in two batches. The first batch contains three dissimilar thought experiments, which are arguably the most controversial cases considered in this thesis. The second batch contains three rather similar thought experiments which have not raised as much attention as the first batch. To begin, consider the first batch of thought experiments:

2.1.1: The Twin-Earth

> Imagine, for instance, that tomorrow God creates a world exactly like ours, such that the only difference is that the people in it—our doppelgangers—and everything else, were created from scratch. Should we say that those people—the people who look exactly like us in twin Africa, Asia, Europe, and so on—constitute races any less than we do, just because the members of each apparently racial population have no

> distinctive ancestries? That seems excessive (Glasgow, A Theory of Race, 2009, p. 32).

2.1.2: George's Appearance-Transformation Machine

> [Imagine] the case of 'George,' who has all-black ancestry but invents a machine to change his appearance so that he looks white. […] It was found that 41 percent of respondents [to a survey study said] "George was not black after his transformation" […] [If] two-fifths of people can deny that George is black, it's hard to see how it's conceptually true that one must have the same race as one's ancestors. (Glasgow, A Theory of Race, 2009, pp. 64-65)

2.1.3: The Dalai Lama

> [Imagine] that some activists […] decide that the best solution [to racism inflicted by the Police] is just to make us all look the same. [They] develop a chemical agent that changes the genetic makeup of anyone who ingests it such that they end up looking exactly, and permanently, like the Dalai Lama. They then infuse the global water supply with this agent, and sure enough, within a few weeks, every human being on [Earth] looks like the Dalai Lama. [If] for a few generations we keep the ancestral populations we had prior to the change [so that while] everyone looks like the Dalai Lama [, then] ancestral populations have not (yet) faded away. [Therefore,] races must, by definition, be visibly distinct, but populations need not be. (Glasgow, Is Race an Illusion or a (Very) Basic Reality?, 2019b, pp. 121-122)

These three examples are possibly the most controversial thought experiments found in this debate, resulting in a highly critical reception. They have elicited important methodological criticisms to consider.

Quayshawn Spencer criticizes Glasgow for the use of “intuition-based thought experiments” (2019b, p. 232). He argues that theorizing about any large enough English-speaking community’s dominant meaning(s) of the word ‘race,’ like that of American English, would be unreliable based on modern statistical theory and have experimentally been demonstrated to be so in studies of American race thinking (p. 237). Therefore, thought experiments of this kind do not help in the metaphysics of race. According to Spencer, intuition-based methods are generally suspicious, especially where they do not take into account lessons from modern statistical theory.

Chike Jeffers criticizes Glasgow for the results drawn from the *Dalai Lama* thought-experiment, whereby the introduction of a chemical agent in the drinking water supply makes everyone look the same (2019b, p. 184). Jeffers argues that a more reasonable reaction would not elicit what Glasgow alleges to be the conclusion, especially since different outcomes could easily be imagined (p. 184). In other words, the thought experiment underdetermines the conclusion. Sometimes thought experiments do not compel participants toward their intended goals. According to Jeffers, Glasgow fails to see that his thought experiment is leaving alternative outcomes open.

Michael Hardimon outright dismisses Glasgow’s *Twin-Earth* and *George’s Appearance-Transformation Machine* thought experiments (2017, pp. 45-47). He argues that “empirical concepts are invariably tied to specific contingent features of the empirical world” and that in “constructing sci-fi examples to test the specification of their content, we should respect the contingencies to which the concepts are tied" (p. 47). For Hardimon, this invariable tie should be a meta-philosophical lesson about the ways that thought experiments fail to establish their intended results. Empirical concepts are contingent but not arbitrary, and thought experiments that fail to take this into account lose their grip on reality. Thus, according to Hardimon, Glasgow's method is flawed in the way it handles empirical concepts.

In reply, Glasgow defends his way of using thought experiments and encourages their use for clarifying the limits of the ordinary concept of race (2019a, p. 263). According to him, imagining fictional scenarios can reveal one's willingness to use a term in hypothetical cases, which provides evidence for what the term could and could not refer to in practice (p. 265). Thus, one can learn about the extent of one's concepts and notice their facility through thought experimentation. Glasgow defends his method and the way he employs thought experiments, as well as their use for analyzing ordinary concepts like 'race.'

As these disagreements make evident, thought experiments are a contentious business. Hardimon, Jeffers and Spencer criticize Glasgow and take his use of thought experiments to be erroneous. I think the number of good criticisms facing Glasgow should be a sign that thought experiments are being misapplied.

The thought experiments considered so far disclose important aberrations to investigate. Yet, there are further methodological issues to consider. To elucidate them, I now turn to a second batch of thought experiments by Glasgow:

2.1.4: The Utopian Babies

> [Imagine] a world of only babies. Everyone else has died off. A new technology keeps the newborns alive and cares for them until they can care for themselves. Before the adults perished, they acted to prevent the terror wrought by centuries of unjust racist behavior. Wanting their children to avoid the same racial struggles with which humanity had plagued itself, the parents decided to wipe any trace of racialization. They destroyed any records that refer to our racially fraught history. In fact, just to be safe, they erased all history and culture other than what was needed to provide the babies with enough science to maximize their well-being. All babies are given equal resources. A variety of therapies become available to allow them the equal chance for equal health outcomes. And so on. Any other information is eradicated in an attempt to present the Reboot Generation with a social blank slate. […] Surely the babies would still have their races after [everyone else] perishes, if they have any races to begin with. (Glasgow, Is Race an Illusion or a (Very) Basic Reality?, 2019b, p. 133)

2.1.5: The Disaster

> Everyone above the age of ten months is being killed by a virus that itself will expire as soon as it kills the last person who is more than ten months old. In a furious effort as they await their doom, the remaining scientists devote themselves to finding a device that can keep the infants alive until

> they are old enough to survive on their own. (Glasgow, A Theory of Race, 2009, p. 121)

2.1.6: Temporary Amnesia

> We are all simultaneously struck by an agent that causes us to forget our systems of racial classification. Any time we start to racially classify ourselves, our cognitive apparatuses short-circuit. One hour later, cognition reverts to its pre-amnesiac state, and racial classification resumes. [In this case,] even though we have no practices of racial classification, it seems counterintuitive to say that we lose our races for the 60 minutes in question. (Glasgow, A Theory of Race, 2009, p. 121)

These three thought experiments have much in common. *The Utopian Babies* case can be seen as the synthesis of certain elements from *The Disaster* and the *Temporary Amnesia* thought experiments. All three have elicited criticisms similar to those levelled at the first batch, but the second batch of thought experiments has not attracted as much attention as the ones considered earlier. Spencer, as perhaps the only one offering relevant criticisms for these thought experiments, argues that intuitive appeals are problematically at work in the *Temporary Amnesia* and the *Utopian Babies* thought experiments (2019b, p. 233). He also adds that Glasgow's overall argument is crucially dependent on intuitive appeals in ways that could not count as evidence. Since Spencer is concerned with Glasgow's argument overall, he does not expand on the possible evidential value of intuitive appeals, and neither does he comment on how thought experiments can become epistemically valuable.

Glasgow's belief that he learns something about the ordinary concept of race when he applies it in unusual circumstances raises a few questions. If a thought experiment is about the ordinary concept of race, which is not normally articulated but learned and adopted in practice and relevant ordinary situations, then how could anyone be sure that irrelevant and extraordinary situations reveal anything useful about the ordinary use of the concept?[1] There are no guarantees that an ordinary concept from ordinary contexts would be applicable in extraordinary contexts. Extraordinary circumstances could indeed reveal something that has gone unnoticed about the concept, but they could also reveal something about extraordinary contexts or about how the concept is contextually sensitive. There are many words or concepts learned in ordinary contexts that could fail in extraordinary ones. If they do fail, such thought experiments will not yield epistemically relevant results. Therefore, beginning with contextualized meaning, one could find that their concepts or words easily misfire out of context. Strangely enough, this appears to be exactly how Glasgow concludes his thought experiments. He seems to beg the question by trying to show how his version of the ordinary concept of race leads him to his theoretical conclusions and undermines the positions of other theorists. Instead, what he needs to accomplish is to show how the results, free from his version of the ordinary concept of race, support his position about the ordinary concept. But, because he already takes his position to be based on 'the ordinary' concept of race, his thought experiments yield the so-called 'ordinary' results

[1] It may seem that similar questions may be levelled at any thought experiment. But, in this case, the question targets a distinct feature of Glasgow's method, namely the focus on *the ordinary concept* of race. As such, the question is not applicable to all thought experiments, but only to those purporting to study ordinary concepts.

that he anticipates. This confusing circularity leaves his thought experiments unconvincing.

2.2: Chike Jeffers

So far, Joshua Glasgow has provided plenty of useful thought experiments for review. But, the questionable use of thought experiments is easily found in other sources. This is why I turn to the works of Chike Jeffers. His thought experiments are, in my opinion, far more sophisticated, carefully written and harder to dismiss than Glasgow's. However, they present new aberrations to consider. For this reason, I believe that they are particularly worth being studied. Here, I present two thought experiments from Jeffers as follows:

2.2.1: The Tight Match

> Imagine if any mismatches between the groups picked out by the study of genetic clustering and the groups known as races in everyday thought and social practice were to disappear in the following remarkable way: people with influence over education policy, media outlets, and other means of knowledge dissemination came to be convinced by Spencer's non-essentialist biological realism and, over the course of a few generations, the identification of the genetic clusters at K = 5 as races became common sense, at least in the United States. In this scenario, we would have not just extensive overlap with some bureaucratic categories but a tight match between a set of biologically real groupings and the ideas of what races there are in all major forms of public discourse. This is, importantly, not impossible. It is part of how social norms work that scientific and

> philosophical ideas can shift and reshape them. Were this to happen, would it not then be the case that biology had become the foundation of race? My answer is no. […] A subsequent shift in popular ideas could result in much less connection to anything of biological significance (for example, grouping together European and East Asian under a broad Eurasian category determined by lightness of skin while splitting off dark-skinned South Asians, etc.). This second shift would not be a move away from race toward something else, but simply a reorganization of racial designations and identifications. The previous connection to something more biologically significant would thus be revealed as inessential, for what is essential to race is that people's looks and lineages as tied to places of origin gain social significance. (Jeffers, Jeffes' Reply to Glasgow, Haslanger, and Spencer, 2019b, pp. 182-183)

2.2.2: The English-Bangladeshi Woman

> Imagine a young woman, born in England to parents from Bangladesh, whose dark brown skin has marked her for her whole life as a minority of foreign origin. What should she make of the idea that it would be accurate to classify her as being of the same race as the majority? Faced with a choice between describing herself in relation to white people as racially different in recognition of how her appearance has generated a particular experience and describing herself as racially the same on the basis of the broadness of "Caucasian" or "Caucasoid" as a category, should she see both options as equally reasonable [descriptions of race]? […] I would

> deem the first option more illuminating. [The second option] conflicts with common sense in a way that is best addressed by giving up the idea that it counts as a description of race [and to rephrase scientific insight into] our development and diversity as a species in other [non-racial] terms. (Jeffers, Jeffes' Reply to Glasgow, Haslanger, and Spencer, 2019a, pp. 43-44)

Both of these thought experiments are concerned with common sense understandings of race, and both aim to dissuade scientific revision of common sense. Yet, there are important differences to consider. The *Tight Match* is focused on how racial concepts are disseminated, while the *English-Bangladeshi Woman* is focused on how racial concepts impact the lives of racial minorities.

Spencer argues that one of the ordinary ways that Americans competently talk about race is based on the United States government's Office of Management and Budget (OMB) classification which is used standardly for many official purposes (college and job applications, birth certificates, mortgage loans and so on) (2019a, pp. 78-83). These OMB racial classes are what Jeffers is referring to as the "bureaucratic categories" in the *Tight Match* thought experiment. Spencer argues that the OMB races are biologically real, while they are not purely objective or independent of human interests, and that something is biologically real if it is an epistemically useful and justified entity within an empirically successful biology (pp. 77, 94-103). He argues that population genetics is a successful epistemic program in biology and it can usefully and justifiably divide the human population into several continental groups (pp. 94-103). These human continental groups are, then, biologically real according to Spencer and they are what Jeffers is quoting above as “the genetic clusters at K=5.” Spencer’s argument concludes that statistically speaking the OMB races line up well enough with these human continental groups to make the bureaucratic racial scheme a practical proxy for biology (pp. 103-104). In a nutshell, this is what Jeffers means by “Spencer’s non-essentialist biological realism” in the *Tight Match* thought experiment.

The *Tight Match* thought experiment asks its participants to do something very simple. It prescribes them to imagine that a common sense concept is reformed to better adhere to a scientific model, just to be misaligned again. In more detail, participants are to imagine that the ordinary concept of race changes from the “extensive overlap with some bureaucratic categories” described by the OMB classification to form “a tight match between a set of biologically real groupings and the ideas of what races there are in all major forms of public discourse” only to lose this “tight” correlation (Jeffers, 2019b, p. 182). This is supposed to convince participants that popular ideas are what ultimately organize racial designations and classifications (p. 183). As such, according to Jeffers, biology cannot be triumphant because of the “prevailing social situation” (p. 182). Therefore, dominant ideas must determine what race means, not anything else.

Spencer has two similar criticisms directed at these thought experiments. The first is about the appeal to intuition, in such a way that questions the manner in which Jeffers concludes the *English-Bangladeshi Woman* and perceives his use of thought experiments as nothing more than a strategy to appeal to intuition (2019b, p. 233). The second is about the method: Spencer outright rejects the use of thought experiments, at least when they are of the “intuition-based” variety and attempts to discover the semantic content of commonly used terms and the scope of applying ordinary concepts (p. 232). The very word ‘intuition’ is a point of contention in philosophy for a variety of reasons which I will not address here. However, it is worth noting that ordinary language is not as indulgent as philosophy in employing the vocabulary of ‘intuition.’ Philosophers are often compelled one way or another about a topic and do not know or cannot say why. For a working definition, I take intuitions to be knowledge claims acquired through unknown (or mysterious) sources. Likewise, I understand “intuition-based” thought experiments to involve a variety of metaphysical, psychological and other contentious or dubious mechanisms.[2] Today, science has developed enough counter-intuitive epistemic techniques to unsettle the privilege of intuition. Nonetheless, even the most scientifically minded thinkers continue to interpret their intuitions to be at least seemingly right about something, whether this be about themselves or the world. Given this whimsical epistemic obscurity, it is at least reasonable to think of an appeal to intuition as a likely indication that a philosopher has run out of arguments (Williamson, 2004, p. 109). This is not to say that no such appeal is justifiable, but that appeals to intuition are as a rule questionable. Similarly, Spencer argues that sampling “the thoughts of a single American English speaker” such as a philosopher cannot be a reliable way to “arrive at any ‘core’ semantic content in the widest shared meaning of ‘race’ among American English speakers” (2019b, pp. 234-235). Modern statistical theory shows how “intuition-based” thought experiments are unreliable (pp. 234-238). There may be a place for “highly contextualized” thought experiments which do not claim to speak for such large and

[2] These could be anything like privately held and inarticulate convictions, gut feelings, instincts, desires, whims, inspirations, the subconscious, the linguistic-sense, a-priori convictions, the memory of the forms, the law within, providence, miracles, guiding spirits, whispers of angels, and so on.

diverse populations (p. 237). However, highly contextual thought experiments are very different from what Jeffers has offered.

The *Tight Match* scenario imagines sociocultural signification as a necessary means to keeping the concept of race in sensible order. But in so doing, it relies on an intuitive appeal to the sway of popular ideas to override the empirical dependence of the concept altogether. As Hardimon points out, empirical concepts could find signification purely from sociocultural grounds, but then they would no longer be the same empirical concepts if they were to lose their empirical grip (2017, pp. 45-47). Consider how tomorrow if, instead of something like the freezing point of fresh water at sea level, zero degrees Celsius is recalibrated at the point that is most popular to call ‘freezing’ on an ongoing basis, the unit would lose touch with what it has been physically measuring and would begin to track the popular opinions of the day. Here, ‘empirical grip’ describes the connection between the unit, or concept, and the empirical phenomena it captures. As such, intuitive moves can release empirical grip.

I think the main issue in both thought experiments should be framed in terms of how people ordinarily think or talk about race in commonsensical ways, in contrast to the specialized ways of thinking or talking about race which adhere to some theory. Common sense does include a range of theoretical, semi-theoretical, pseudo-theoretical and unreflective ways of dealing with race. Controversial, sensitive or convoluted topics could easily become muddled where multiple incompatible norms compete, such as when different communities laying claim to common sense do not necessarily hold shared beliefs and value commitments to safeguard controversial topics like race. This is why individuals often adjust their conceptual vocabulary to best align with whatever suits their social circumstances. For example, it is popular in Canada to replace the word ‘race’ with its associates, such as ‘ethnicity,’ ‘background,’ ‘culture’ or other related terms in vogue. However, replacement words still allow racial concepts to continue to work through semantic implication or euphemism by masking the controversial façade of racial concepts and overloading the connotative values of their substitutes. Thus, when inquiring about race, Canadians are more lenient with ‘ethnic’ language, even when they are only satisfied with racial answers. Be this as it may, concern over linguistic piety is not the goal of the thought experiments considered. Instead, they are meant to reveal stable conceptual structures supporting the linguistic façade, what is enabled or confounded, as well as the conceptual tension experienced by their contestation.

To best capture the issue, instead of what counts for ‘common sense’ when it comes to ‘race’ many theorists have opted to seek its ‘ordinary’ usage. The interest in the ‘ordinary’ use of racial terms is closely tied to what is called the ‘operative’ concept/meaning of terms (Haslanger & Saul, 2006, pp. 97-99). Operative concepts have explicit social content because they only conform to social standards in their application (Haslanger, 1995, p. 102). They have the advantage of capturing practical distinctions, even when drawn unwittingly (2012c, pp. 388-389). The fact that humans ordinarily draw consistent distinctions in their conceptual/linguistic practices without being able to give justifying reasons for doing so opens up debate over the right way to articulate the ordinary concept/meaning operative in practice, which is exactly why someone like Hardimon begins his inquiry into race by arguing for ‘the ordinary concept of race’ (2003). For him, the focus is on the “ordinary uses of the English word ‘race’ and its cognates” (2017, p. 27). When it comes to Glasgow, the focus is on “how the term ‘race’ is used by linguistically competent ordinary users in the contemporary United States, to describe groups of humans" with an emphasis on whichever concept of race has the most currency in ordinary usage (2019b, pp. 115-116). Likewise, Spencer focuses on a dominant United States-based "ordinary race talk" called the OMB classification, because it is commonly "used in ordinary discourse" (p. 81). The target in each case is the 'ordinary' usage rather than the 'common sense' understandings of race targeted by Jeffers.

In contrast to the 'ordinary' usage, 'common sense' casts a wider net. It also extends over what is called the 'manifest' concept/meaning that captures what an agent takes themselves to be applying or what they are attempting to apply in a particular case (Haslanger & Saul, 2006, p. 99). These differences may seem subtle, but they could make an important difference in thought experiments.

In this debate, 'ordinary' usage, 'operative' concepts/meanings and 'common sense' notions more or less perform the same role, albeit 'common sense' is not always about the 'ordinary' usage of or the 'operative' concepts/meaning of 'race' as observed in linguistic practices. So, as a word of warning, the reader should keep in mind that different theorists in this debate tend to use different technical terms, something which could make the debate hard to follow.

Chapter 3: Metaphysics of Race

In this chapter, I clarify what is meant by ‘metaphysics’ in the metaphysics of race debate. These remarks are rudimentary, and yet they help to clarify the distinct theoretical commitments of each camp of the debate considered. ‘Metaphysics’ and ‘Ontology’ are labels describing a certain field of study in philosophy or a certain group of (meta)theoretical questions broadly speaking. But this chapter is not intended to clarify metaphysics, but to clarify things to better understand the metaphysics of race debate. These terms are used in various ways depending on the tradition and school of thought employing them, making any strict definition controversial. Fortunately, I only seek working definitions for my purposes in this thesis. These definitions help articulate the diversity of positions on offer and the issues debated.

What do I mean by Metaphysics? Peter van Inwagen claims that the best definition of metaphysics he has come across is one he was introduced to as an undergraduate: “metaphysics is the study of ultimate reality” (2018, p. 1). Still, I take it that a simpler definition would drop the word “ultimate” and describe metaphysics as the study of “that which is real, insofar as it is real” (Wolff, 2006, p. 1244). Metaphysical questions are sometimes believed to be ‘ultimate’ in grandiosity. But it would be more accurate to say that they are ‘ultimate’ questions because they are those questions which ‘ultimately’ arise in the sense of arising inevitably. Unfortunately, metaphysics has a bad reputation for grandiosity,[3] but I do not view metaphysical issues as grand. My working

[3] *The Concise Encyclopedia of Western Philosophy* (2005) opens the entry on metaphysics with the following: “Metaphysics is that part of philosophy which has the greatest pretensions and is exposed to the greatest suspicions. Having the avowed aim of arriving at profound truths about everything, it is sometimes held to result only in obscure nonsense about nothing” (Strawson, p. 242).

definition of ‘Metaphysics’ is this: *the examination of reality in terms of its most basic and most general categories, concepts, notions or theoretical presuppositions*.

What do I mean by Ontology? It is common practice to use ‘Ontology’ as a synonym for ‘Metaphysics’ (Butchvarov, 2015, p. 662). This is because they are inseparable, the former being the core of the latter. Ontology is the metaphysical study of, or an account of, what there is[4] in reality, and tends to be used in a deflationary way. The difficulty is that specifying what constitutes reality calls for explaining what those constituents are, how they hang together,[5] and why other competing accounts of these things are wrongheaded—which is itself a metaphysical project. In my opinion, it is not entirely possible to do one without the other. Yet, modern philosophical movements tend to be averse to the speculative character of metaphysical systems, often considering them a waste of time, while accepting ontology as a worthier attempt to study and debate various inevitable theoretical commitments.

One source for this aversion can be found in a method that purports to solve metaphysical and epistemological problems by eliminating ontological commitments deemed unnecessary. In an appeal to the principle of parsimony, sentences that seem committed to unwanted entities are rephrased (i.e., paraphrased), and regimented in symbolic logic. This method is associated with two developments in philosophy. Firstly, Bertrand Russell’s theory of descriptions (1905) shows how ontologically misleading names (i.e., proper names without bearers) in some sentences can be eliminated if those sentences are rephrased to replace the names in question with their functionally

[4] I borrow “what there is” from W. V. Quine’s paper “On What There Is” (1948).

[5] The phrase ‘how they hang together’ is inspired by Wilfrid Sellars’ turn of phrase in “Philosophy and the Scientific Image of Man” (1963).

equivalent logical constructions to escape unwanted ontological commitments. Secondly, W. V. Quine's doctrine of 'ontological relativity'[6] (1969b) together with his criterion of 'ontological commitment'[7] (1948) have normalized hypostasis[8] by perpetuating the appeal to the pragmatic virtues of theory. The adoption of these and other deflationary methods has helped to make the misalignment of language use with its purposes the focus of ontology.

For my part, I do not believe that ontology is lodged in language or any particular system of logic. Ontological commitments are observable beyond language, being manifest in perception and practice. To be more accommodating, I take a more tolerant attitude to what constitutes reality. So, my working definition of 'Ontology' is this: *the study of what constitutes existence and what counts as being an entity or a kind of entity.*

Debates in metaphysics and ontology are ultimately interpretive preoccupations, involving how something is described and what conceptual consequences such a description brings. But whatever the options may be, some interpretation is needed to enable one position at the expense of others. The job is twofold: to organize reality (classify/categorize) and to simultaneously revise the criteria used for organization (define/schematize).

3.1: Nomenclature

In this section, I go over some of the most elementary nomenclature in metaphysics to clarify important distinctions between the positions of the metaphysics of

[6] Quine's doctrine of 'ontological relativity' holds "it makes no sense to say what the objects of a theory are, beyond saying how to interpret or reinterpret that theory in another" (1969b, p. 50).
[7] Quine's criterion of 'ontological commitment' states "a theory is committed to those and only those entities to which the bound variables of the theory must be capable of referring in order that the affirmations made in the theory be true" (1948, p. 33).
[8] Hypostasis describes "the acts of positing objects of a certain sort for the purposes of one's theory" (Delaney, 2015, p. 487).

race considered in the next chapter. Take for example the terms ‘class’ and ‘category.’ It is important to note that they are not always interchangeable. On the other hand, metaphysicians are fond of using the term ‘kind’ for a metaphysically exceptional ‘type’ of existent in the world. However, a lack of consensus over the metaphysical implications of such terminology, their technical variability, and the closeness of their roles to non-technical everyday colloquialisms of English (or all living languages) undermines their clarity. Like many other words in philosophy and the sciences, these terms have multiple discipline-specific uses which may not align well with their non-technical usage. In the following, I narrow my attention to drawing the terminological differences which I find salient for this thesis.

The goal of this exercise is this: theorists in the metaphysics of race work with the same terms in very different ways, which makes the debate hard to decipher. Once I specify what a term does and does not mean in metaphysics for the purposes of this thesis, I evaluate the arguments in metaphysics of race independently of the theorist’s choice of language. This allows for a sharper comparative analysis.

3.1.1: Type

What is a type? I understand ‘type’ as *a simple or compound characteristic identifiable in different characters or token items*. This is my way of describing the type-token distinction[9] in an easy-to-grasp, yet accurate, manner. In more detail, a type is a representational form (e.g., the sign for ‘A’) or a representational object (e.g., the typeface or stamp for ‘A’) identified by its token objects (e.g., every occurrence of the letter ‘A’) (Rastier, 2004, p. 434). A token is said to exemplify or instantiate (i.e., typify)

[9] The type-token distinction was first introduced by Charles Sanders Peirce ([1906-8]1998, pp. 480-488).

a type by virtue of possessing the type's characteristic features (Bach, 2015, p. 1089). Types are usually[10] employed in classification projects, usually to generalize from particulars. This much I take to be a useful working definition.

Unfortunately, the type-token distinction gets used as a proxy for the universal-particular distinction in metaphysics (Guter, 2010, pp. 208-209). Types do share at least some of the characteristics of universals, in having instances or being repeatable, abstract, predicable and so on (Wetzel, 2018, sec. 3). Also, one can get away with the interchangeable use of 'category,' 'type' and 'universal' in many occasions. For example, 'redness' or 'fluidity' could be labelled categories, types or universals. Such a troubling degree of terminological promiscuity warns not only that metaphysical conventions are confusing, but that they actively invite equivocation.

3.1.2: Class

What is a class? A class is any consistent list of items that share a common feature, often used synonymously with 'set' which is any consistent list of items whatsoever (Iannone, 2001, p. 110). In general, any group could be called a class or a set. But importantly, a class is understood as the extension of a concept (Maddy, 2015, p. 166). Conversely, a concept itself can be understood as a principle, a mental representation or an ability that guides classification (Butchvarov, 2015, p. 194). So, *classes are sets generated by concepts (expressed by predicates).*

[10] 'Type' could be understood in terms of the logical theory of types, pioneered by Bertrand Russell in *Principia Mathematica* (1910), whereby types are regimented levels/orders of classes/sets in a set-theoretic hierarchy where class/set membership is only open to the level/order below, and no class/set can be a member of itself (Menzel, 2015, p. 1087). Put differently, a 'type' is any given class or set that is the extension of a predicate, but does not apply that predicate to itself (Ayer, 2005, p. 338). However, the logical theory of types explains a contextual use of the term in metaphysics, since taking types to be levels/orders of classes/sets, or classes or sets, does not account for how types could be established on the basis of prototypes to make self-predication possible.

3.1.3: Category

What is a category? This is difficult to explain because the term is used vaguely on a regular basis and has multiple possible meanings[11] in philosophy, symptomatic of the metaphysical role assigned to it in different traditions. Roughly speaking, categories are both ultimate classes—the most general system of classifying entities—and/or the ultimate concepts constituting those very ultimate classes—the most general and abstract systems of ideas necessary to understand reality (Meiland, 2015, pp. 146-147). The latter identifies a category according to its conceptual value (e.g., the colour red), and the former according to the class that falls under the concept shaping the category (e.g., every instance of the colour red). Perhaps it is best to describe categorization as the act of classification that breaks wholes into parts for theoretical purposes, instead of bundling individuals into aggregates. In other words, if classes are unified under conceptual similarity, categories are unified according to their conceptual differences organized in theory. Thus, *categories are classes generated by theories*.

In this thesis, I take ‘category’ to be the most elementary distinction made in a metaphysical system. Given that a category is only justified in a system of categories (also known as a conceptual scheme), I take it that any criticism made against a particular category is going to be metaphysically loaded against a system of categories. Moreover, the role played by the term ‘category’ is defined by the metaphysical tradition employing it. So, avoiding metaphysical import often means avoiding the use of the term. For instance, one could simply describe a category as an “ultimate type” to capture what

[11] Category could mean predicate-types (Aristotle, Categories, 1963, p. 5 [1b25]), logical functions of judgement (Kant, Critique of Pure Reason, 1998, p. 206 [A70/B95], 212 [A80/B106]), or the most fundamental divisions of our systematization of the world (Westerhoff, 2005, p. 135), among other things.

philosophers call a ‘difference in kind’ (Ryle, 2005, p. 73). But that could miss their systemic embeddedness since types do not carry the same theoretical implication.

3.1.4: Kind

What is a kind? It is not easy[12] to say. Overall, the tradition in metaphysics and philosophy of science has it that a kind is a unique[13] category of real entities with modal implications for its members (Harper, 2015, p. 701). For example, if Socrates (an individual entity) is a member of humankind (a category of natural entities), then Socrates is necessarily a human being. Realism about kinds is, loosely, the position that some predicates are based on mind-independent realities in the world. This position holds that kinds must be understood as ‘real kinds’ because their modal necessity is non-arbitrary. Moreover, something like this is often called ‘natural kinds’ since unnatural, gerrymandered, arbitrary, stipulated, conventional and non-existent predication is also conceivable. So, (*real/natural) ‘kinds’ are categories of (mind-independent) real entity types*.

Before going any further, I should make my position clear. I subscribe to Ian Hacking’s call to “[delete] every mention of natural kinds” in metaphysics and the sciences (2007, p. 229). But, since the terminology of ‘kinds’ is almost inescapable today, I understand the notion based on the unorthodox assumption that underpins its usage in the contemporary discourse of social ontology: categories dividing nature into classes of

[12] ‘Kind’ is difficult to explain because it is a colloquialism turned buzz-word within the Anglo-American philosophical tradition. It should be noted that the tradition of ‘kinds’ only exists in the English language, inherited from the works of William Whewell and John Stuart Mill (Hacking, 2007, pp. 214-215, 224). This also goes for ‘natural kinds’ which is another way of saying the same thing, but coined later by John Venn, famous for the Venn diagram (Hacking, 1991, p. 110).

[13] Ian Hacking captures this uniqueness: “There is a unique best taxonomy in terms of natural kinds, that represents nature as it is, and reflects the network of causal laws. We do not have nor could we have a final taxonomy of anything, but any objective classification is right or wrong according as [*sic*] it captures part of the structure of the one true taxonomy of the universe” (1991, p. 111).

(often predicate) types are coextensive with natural kinds, such that natural categories and natural kinds are two sides of the same theoretical coin. The reality of ‘kinds’ pertains to the world while ‘categories’ pertain to language and theory as concepts of kinds (Khalidi, 2015, p. 97). To use a cartographic analogy, a metaphysical system categorizing nature draws the map and its coextensive natural kinds name territories.

In a broader sense, natural kinds are ancient ideas. Philosophers fondly recall the Phaedrus’ metaphor of carving nature at the joints, suggesting that nature is conveniently structured to serve human needs (Plato, 1995, 265e). This view about nature is no longer popular, but not because of a lack of delineating joints in nature, since there are actual joints in nature, namely, the anatomical articulations of skeletal structures commonly known as ‘joints.’ Instead, there has been a general shift towards accepting fuzzy, overlapping and gradient aspects of nature as well as acknowledging the limits of human discernment.

With these simple admissions, it seems unreasonable to assume that the linguistic norms privileging certain natural predicate types as ‘kinds’ should be ontologically substantial because of their dependence on language and mind. But, since Nelson Goodman’s (1955) “New Riddle of Induction” has alarmed naturalistically inclined philosophers about the possible choice of predication in describing scientific observations, preventing the use of unnatural, relative or conventional predicates in inductively valid scientific hypotheses has become a priority for many (p. 72). Some inductive arguments are extremely reliable because they make conclusions about things in nature that are invariant. The argument could be premised on small samples, or even one single sample, of invariant populations/materials. The success of these inferences is not due to their formal structure, but because the samples they rely on are natural kind samples “whose members do not vary that much in the properties they exemplify” (Godfrey-Smith, 2003, p. 585). If the choice of predication is limited to natural kind predicates, such inductive arguments, called ‘projection,’ remain inferentially reliable (p. 583). But if alternative predicates describe the same observation, the inference loses its reliability. In response, some have tried to justify the eligibility of select natural kind predicates by appealing to the success of scientific theories, instead of directly appealing to nature’s inherent structure. In practice, however, when philosophers are convinced that a predicate is not gerrymandered and possibly suitable for use in scientific hypotheses, they are likely to call it a ‘natural kind.’

There are two live approaches to natural kinds today. The first approach is epistemic and identifies a grouping as a natural kind when it plays an important role in scientific inquiry. This approach has gained prominence since Richard Boyd’s (1989) homeostatic property cluster theory (HPC) which identifies natural kinds in correspondence with loose but stable collections of properties (i.e., a cluster) possessed by typical kind members (pp. 11-18). Membership in a kind is not a matter of possessing all and only a certain requisite set of properties, but of possessing enough of the properties of the cluster held in equilibrium or homeostasis by a causal mechanism to be able to serve in successful scientific induction, prediction, projection or explanation (Khalidi, 2023, pp. 20-21). In short, a natural kind is an identifiable stable property cluster that is causally linked to other properties in nature.

The HPC theory seems uniquely suited to handle fuzzy cases. However, many examples of natural kinds do not easily fit in the loose and stable cluster mould. For example, electrons are the kinds of things that cannot lose their charge. Their charge does not vary nor is it a contingent property to be considered one part of a cluster of properties.

The HPC theory appears to confirm many things as natural kinds as long as they correspond to some scientifically explainable property cluster. This means that kinds can be confirmed as natural only in the shadow of their success in scientific reasoning, which seems oddly similar to scientific conventionalism[14] about natural kinds. But, if natural kinds owe their grip on reality to scientific theories, there seems to be nothing unique about natural kinds in metaphysics. If so, HPC natural kinds are theory relative.

The second approach to natural kinds is rooted in the revival of older philosophical ideas[15] within philosophy of language. Saul Kripke argues that common

[14] For example, Goodman's "comparative entrenchment" view finds valid predication to be based on successful linguistic practices in science (1955, pp. 108, 121).

[15] The notion of 'natural kinds' is closely tied-up with 'category' in philosophy. For Aristotle, categories are the most basic and general predicate-types possible (1963, p. 5 [1b25]). Correctly identifying what something is yields the category of substance (a non-reducible fundamental source of being for Aristotle) (p. 75). A subject-expression can specify an individual object (e.g., Socrates) as a primary substance or, conversely, a predicate-expression can identify the species or the genus that an individual belongs to (e.g., human or animal) as a secondary substance (p. 79). Following Aristotle, Medieval philosophers considered species and genera to be what is today called natural kind substances (Iannone, 2001, p. 144). There is more to secondary substances, but this is enough to show that these metaphysical notions do not easily come apart. The language of kinds is hopelessly loaded. The Aristotelian tradition encourages the metaphysical link between 'category' and 'kind.' But 'type' and 'universal' are also entangled. Indeed, triangulating the terms 'category,' 'universal' and 'type' surveys the region reserved for 'kinds' in metaphysics. For example, kinds could be conceived as universals whose particular instances are objects and, likewise, natural kinds could be conceived as universals whose particular instances possess properties as a matter of natural necessity (Lowe, 1997, pp. 36, 40). Conversely, kinds could also be understood in terms of types (Wetzel, 2018, sec. 4.1.2). According to Hilary Putnam, one standard way of informing someone about the meaning of a natural kind like 'water' is to use a "stereotype—a standard description of features of a kind that are typical" ways of marking out 'water' from other kinds of things (1975, p. 147). In other words, I can inform someone of what water is by describing its "type" or "normal form" which is colourless, transparent, tasteless, thirst-quenching and so on (pp. 190-191). Others have conceived of kinds as having their membership modelled on the tokens of a type-forming category (i.e., an archetype) (Bromberger, 1992, pp. 176, 201). In this sense, water is not the type of thing described by a stereotype, but, instead, certain types of things are identified as water, including the type of thing described by the stereotype. Furthermore, it has been argued that universals are kinds, where kind and type are taken to be identical (Moreland, 2001, pp. 75-80). At this point, it must be clear that these terms mean whatever metaphysicians want them to mean, and the language of natural kinds is a hopelessly entangled sticky web.

names designating natural kinds are rigid designators[16] (i.e., names that designate the objects they refer to in all possible worlds), that reference to kinds is made by the ostensive identification of their samples in a causal chain which could, with the scientific discovery of their natural structures, reveal their necessary theoretical essences (1980, pp. 48, 127, 134-140). Hilary Putnam makes the same point in pushing for an externalist theory of reference and argues that the reference of a natural kind is fixed indexically by identifying typical likeness relations (1975, pp. 147, 152, 190). He also holds that reference fixing rests on a certain division of linguistic labour, whereby a non-superficial criterion of likeness is determined by the judgement of the right experts[17] (pp. 144-146). In this tradition, natural kinds have theoretical essences, fixed via causal and social practices.

[16] It is peculiar to suggest that the lax role of common names for natural kinds could be strict enough to serve the demands of rigid designators. Natural kind language is well-suited for establishing communication rather than theoretical rigour, and it seems incontrovertible that mature scientific theories develop special terms for analysis in place of natural kinds (Quine, 1969a, pp. 121, 137-138).

[17] The connection between natural kinds, essences and experts is based on Putnam's sociolinguistic thesis: an empirical "hypothesis of the universality of the division of linguistic labor" (Putnam, 1975, pp. 145-146). Strangely enough, Putnam does not say whether this thesis is confirmed in linguistics in the way that he presents it. I do not doubt that such a division of linguistic labour is observable, however, I question its characterization. Admittedly, the division of labour in language is an easily observable fact. That is why dictionaries exist. In one occasion, Putnam analogizes words to tools "like a steamship which require the cooperative activity of a number of persons" (p. 146). But there is a clear misalignment in this analogy, because a steamship could be operated without any of the rank-and-file specializing beyond their limited roles. In such a case every role has its specialist and there are no experts for the entire ship. I suspect that much of our vocabulary, including natural kinds, are in such a predicament. When it comes to 'water,' for instance, who is the expert? The chemist, the geologist, the environmental scientist, the water sanitation engineer or the water sommelier? It depends on what one means by 'water' in each case. If I want to ask about the molecular structure of water, I have to consult a chemist. But if I want to ask about the comparative taste of drinking water from various sources, I have to seek a sommelier. And this criticism also extends to Kripke's natural kind essentialism. With hard-water it is not H_2O that tastes chalky, but the other minerals present in the liquid. And, yet, it is perfectly acceptable to associate the chalky taste to the water sample itself, because hard-water is a kind of water. Natural kinds do not have experts. However, there are experts who specialize in related topics (e.g., a scientist, an engineer or a sommelier specializing in water).

So far, I have surveyed the dominant ways of thinking about key metaphysical concepts. But even though I try to remain as neutral as I can, in my application of these concepts I cannot remain completely neutral. So, here, I take a stance.

In the Kripke-Putnam approach, I do not see the benefit of framing certain necessary identity relations that hold between natural kinds and scientific theories as ‘essential’ besides fulfilling the needs of the essentialist. The identity relations hold between the object or substance (metaphysical) and its description under a theory (epistemological), rather than the object or substance (metaphysical) and its essential nature regardless of the theory (metaphysical). What establishes the identity relation is the scientific theory, not the description. Consider how, in chemistry, the theoretical category ‘water’ includes non-H_2O descriptions composed of different isotopes of hydrogen and oxygen, such as heavy water (D_2O) or light water (DDW). Ordinary drinking water, in its pure form, is identified as H_2O because it is not coextensive with the use of ‘water’ in chemistry. But if there are different referents for the common noun ‘water’ in chemistry, the noun is semantically contextual and not a rigid designator. If so, natural kinds may have no essences, only theoretical identities.

The doctrine of ‘kinds’ remains highly controversial. The appointment of certain predicable common nouns as metaphysically insightful seems unscientific and unnecessary. Natural kinds are the by-product of ontological, epistemic and semantic theories being entangled, and I believe the semantic theory is the weakest link for overgeneralizing theoretically contextual uses of natural kind common nouns to the rest of language. Because of this, the HPC approach seems more convincing, though not without controversy.

In this section, among the terms covered, ‘class’ and ‘set’ have been the easiest and least contentious to unpack. Then, there is ‘type’ which seems to be generally easy to understand. The term ‘category’ also lends itself to a workable definition, as the most basic and general partition in metaphysics. On the other hand, ‘natural kind’ has been the most convoluted. I believe that ‘kinds’ add no technical value to a metaphysical discussion, other than to encapsulate other metaphysical commitments. So, in line with my own metaphysical (non)commitments, I believe the loose and generic use of ‘kinds’ to be more sensible than any of its technical mutations.

3.2: Social Reality

3.2.1: Social Metaphysics or Social Ontology

What is social metaphysics or social ontology? The field examines social reality in terms of its most basic and most general categories, concepts, notions or theoretical presuppositions. This study (henceforth simply ‘social ontology’) is also concerned with the relationship between human minds and society. For this reason, it involves the philosophy of mind, action, and social science.

3.2.2: Metaphysics of Race

What is the metaphysics of race? Charles Mills characterizes the metaphysics of race as one more theoretical attempt at understanding social reality:

> [There] are basic existents that constitute the social world, and that should be central to theorizing about it. Thus, one readily understands what it means to say that the social ontology of the classic contractarian is an ontology of atomic individuals; that for Karl Marx, it was classes defined by their relation to the means of production; and that for radical feminists,

> it is the two sexes. […] But systemic racial privilege has been an undeniable (though often denied) fact in recent global history, and exploring an ontology of race will contribute to (though not exhaust) our understanding of social dynamics. (Mills, But What Are You Really?, 1998, p. 44)

In following this, the central question for an ontology of race should be: does race constitute the social world? Whether or not race is a constituent, I take the upshot of this view to be a clear metaphysical question: *what is race, really?*

3.2.3: Social Kind

One of the central debates in the metaphysics of race revolves around the claim that the concept somehow uniquely describes a social phenomenon which may not have any physical or biological basis in nature. The claim is that racial distinctions are grounded[18] in social practices, not natural processes, and for this reason, they rest on a non-racial reality and are not fundamentally immutable. In other words, the issue is whether race is best described as a 'social kind' instead of a natural kind.

[18] 'Grounding' is a technique for giving a metaphysical explanation (usually across ultimate metaphysical categories). There are different theories of grounding, but generally it determines the condition of existence for some entity from a *less fundamental* ontological level/*dependent* category of being by identifying a fact from a *more fundamental* ontological level/*independent* category of being, through a non-causal asymmetric relationship (often articulated in terms of *arising* from, being *constituted* by, being *constructed* as, *generating* out of, *consisting* in, existing *in virtue of*, or simply coming to be *because* of some underlying basis) (Trogdon, 2015, pp. 430-433). For example, the fact that sunshine gets obstructed by my body grounds the fact that I cast my shadow on nearby surfaces. This ontological dependence explains the status of the shadow as an object. Similarly, many social facts are claimed to be grounded in other more fundamental/independent facts about reality to justify their special theoretical treatment. In the 'Standard Model,' popularized by John Searle, the 'constitutive' rules tend to have two jobs: as *constitutive*, they assign (i.e., constitute or generate) a status to an object and as *regulative* they describe that object's normative (or deontic) conditions (Searle, 1969, pp. 33-42). This is because institutional facts describe status functions (i.e., roles) that imply normative powers (2010, p. 23). G*rounding* is, thus, accomplished by laying out the status and normative conditions specified by a constitutive rule.

What are social kinds? Social kinds are a variant of natural kinds in social ontology. In a nutshell, *social kinds are categories of real social entity types*. In terms of the HPC, social kinds are likely to be stable property clusters linked to causal mechanisms in social reality that, if successfully detected, are described by the social sciences. In terms of the Kripke-Putnam view, social kinds are likely to have their references fixed by speech acts or mental states. They are the variety of ontological categories that share a dependence on human social life (i.e., minds, institutions or social practices). Unlike natural kinds, they are believed to be dependent on social structures[19] over and above natural reality.

3.2.4: Social Construction

What makes a category of social fact types a social kind? The answer depends on the social model used. A popular model in social ontology, sometimes called 'the Standard Model,' is based on the works of John Searle. It holds social facts to be the products of collective intentionality—the intentional attitudes and actions of individuals jointly directed at objects (Hakli & Mäkelä, 2015, p. 999)—imposing social functions on objects. It also takes institutional facts, a subset of social facts, to consist of the collective acceptance or recognition of the assignment of social statuses with social function (Searle, 1995, pp. 26, 41, 124). Some social rules establish status functions (i.e., roles) and can be expressed in the form "X counts as Y in context C" where "X" is a physical or lower order fact and "counts as Y in context C" is a "constitutive rule" describing the

[19] Some hold that social structures are merely phenomenal. For example, according to Iris Marion Young (2011) "[social] structures are not a part of society" but, instead, "become visible [by] a certain way of looking at the whole of society [that sees the] patterns in relations among people and the positions they occupy relative to one another" (p. 70). In other words, social structures are perspectival in nature as aspects of the same non-social ontology from enough of a distance to the personal point of view.

institutional fact "Y" in the appropriate social context "C" where "X" qualifies as "Y" (p. 28). To use Searle's favourite example, American dollar bills gain the institutional function of having monetary value when they become, as a rule, widely recognized to function as currency in the economy: "Bills issued by the Bureau of Engraving and Printing (X) count as money (Y) in the United States (C)" (p. 28). Dollar bills are endowed with economic status because enough minds have been jointly directed to accept or recognize their constitutive rule.[20] Therefore, a type[21] of object plays the role of money because 'money' is a social category of entities/a social kind-concept widely recognized to have the right function in the economy.

3.2.5: The Social

What makes anything social? Whether something is considered to be social depends on the way 'the social' is categorized in theory. One way[22] to differentiate the

[20] It is worth acknowledging that not all institutional roles are constituted by rules. For instance, Max Weber, argues that "traditional authority" is constituted by "precedents and earlier decisions" rather than "rules" ([1921]1978, p. 227). Many traditional social statuses are indexically defined by way of exemplary tokens, for instance, a traditional ruler refers to a person with the same particular status as the paradigm case of a particular ruler experienced by members of a traditional community or described in the community's traditional forms of knowledge (Rust, 2021, p. 322). Such roles are grounded in the traits and behaviours of preceding exemplars of that role, rather than rules describing their constituting conditions (p. 322). Still, traditional roles are institutional kinds.

[21] Not all social roles are assigned to types of things. Tokens can also be social kinds. For example, in Canada, the status of a prime minister (a category/kind-concept) is assigned to one individual person (one token) at a time. The leader of the political party with the largest number of votes in a federal election (X) counts as the prime minister (Y) in Canada (C). So, the kind 'the current prime minister of Canada' only extends over the token individual who in fact holds the title at a given time. This shows how social kinds are not always social fact types, but sometimes social fact tokens.

[22] Another way to categorize the difference is by distinguishing the natural from the artificial, where only the artificial is identified as the social. For example, sugar is understood to be natural regardless of its refinement process, because it is chemically available in sugar cane and other sources, but sucralose (a common artificial sweetener) is taken to be artificial because it is chemically synthesized through a patented process. It is generally assumed that human made products are artificial as opposed to those that are naturally occurring. However, such a distinction is anthropocentric. Some human crafted things easily fall into the natural category, such as laboratory produced diamonds and technetium, domesticated animals and many varieties of artificially selected, crossbred or genetically modified or even engineered crops. Conversely, some of the things that fall under the natural category are the artifacts of non-human animals, such as bird nests or beaver dams. Furthermore, some social animals build highly complex communal architecture, such as nests constructed by bees, ants and termites. In these cases, not only are artificially crafted things natural, but at least in the case of social animals, social reality is itself natural. Since

social is by claiming that it is uniquely grounded in the mental activities shaping the social world. Consider how it is generally accepted to claim that bees and ants are creatures with highly sophisticated hierarchical societies, but it seems that their social attributes and enterprise are not mind-dependent in a comparable way to human societies. The rule that a bee or an ant counts as a queen for its colony does not rest on the acceptance of certain rules by colony members. It rests on their conformity to the rule. The life of these creatures is as natural as it is social, and the mind is not a special substance to bestow social kinds with a unique ontological status. A better approach would seek the underlying structures of social reality without relying on mental substance to mark the social. Delineating the social remains ontologically ambiguous.[23]

Even as mind-dependent, some social kinds are not concept-dependent as described in Searle's Standard Model. For instance, large-scale social facts like recessions, identified by (statistical) generalizations over smaller social facts, do not depend on any intentional states regarding their concept (Thomasson, 2003, p. 606). Consider, also, how some facts about racism seem to depend on the presence of certain racist beliefs, but they do not depend on the existence of beliefs about racism itself (p. 606). Similarly, explicit or implicit beliefs and intentional attitudes about racism or

everything in the world does not revolve around humanity, there is no reason to think that human contamination marks the difference between artificial and natural. Short of this, the artificial fails to be the basis for differentiating the social.

[23] The dichotomy between a mind-dependent social reality and a mind-independent nature can be misleading. In contrast, a well-established school of thought in the social sciences outright excludes the psychological and the biological from its conception of the social. For example, in *Method* (1982), Emile Durkheim argues that social facts consist in "manners of acting, thinking and feeling external to the individual, which are invested with a coercive power by virtue of which they exercise control over him. Consequently, since they consist of representations and actions, they cannot be confused with organic phenomena, nor with psychical phenomena, which have no existence save in and through the individual consciousness. Thus [,] they constitute a new species and to them must be exclusively assigned the term social" (p. 52).

recession are not necessary for racism or recession to take place. All agents need to tacitly accept strategies that enable, cause or perpetuate economic recession or racism. But being concept-independent in such a tacit way also means that racism or recession could exist without having to be conceptually or mentally represented for the minds in question (p. 607). So, I ask: is there an ontological difference between such social facts not being mentally represented and the social fact of termite colonies to individual termites? My response is negative, because in each case the relevant concepts are opaque to the relevant social players, whether or not they could in principle be made transparent.[24]

Determining what should count as ontologically social is as controversial as it is difficult. It is very important to keep in mind that metaphysical debates about social reality tend not to be debates about the choice of theory in the social sciences, but emphatically about the shared social world itself. How social kinds are to be interpreted

[24] In the social sciences, some stable and recurring regularities, like many tacit dispositional trends and higher-order facts, are explained by underlying causal mechanisms, such as unconscious, ideological or structural processes. One popular approach based on Pierre Bourdieu's 'theory of practice' asks: "how can behaviour be regulated without being the product of obedience to rules?" (1990b, p. 65). Bourdieu's answer rests on a crucial distinction between *rules* for compliance (consciously accessible and explicable at a personal level) and dispositional *strategies* (consciously inaccessible and inexplicable at a personal level) (1990a, pp. 145-160). Strategies correspond to the tendencies of behaviour which arise in the regularities of circumstances where social agents interact. These 'practical circumstances' are the space of social differences where social reality exists outside the mind of agents (Bourdieu & Wacquant, 1992, p. 127). The mind-independent reality of fields can be used to ground dispositional strategies. In contrast, John Searle's institutional 'Standard Model' only relies on mind-dependent rules to ground institutional facts, though he acknowledges that some similar facts do regularly occur "without constitutive rules and without requiring rules" (1969, p. 40). Still, his model falls short of addressing them.

Bourdieu's theory of practice reframes the assumption that social reality always needs to be thought in terms of an asymmetric relation of mind-dependence. His model describes a mutual relation of "ontological complicity" reflected between individual social players and their social environment (Bourdieu, [1981] 2014, p. 306). The dependence relation goes both ways, such that the structure of social reality is in some ways mind-dependent, but the structure of the mind also depends on social structures in important ways. The external structures of practical circumstances or fields (e.g., habits, statuses, positions, institutions) coincide with the internal structures of agency involved in those practices (e.g., 'habitus,' conduct, dispositions, compliance) as mutually constituting and reproducing the same social reality (Grenfell, 2014, pp. 44-45). In this respect, grounding could sometimes go from practices to circumstances, or from circumstances to practices, according to the relation of ontological complicity.

remains debatable. Arguing about the ontological status of social facts in terms of social kinds invites theorists to talk past each other. For this reason, the concept of ‘social kinds,’ as a subcategory of ‘natural kinds,’ is highly controversial. It rests on assumptions about the metaphysics of social structures and is often wielded to secure an independent ontological realm for social entities. In my view, the ontological difference between the natural and the social is not so stark as to grant such assumptions. Thus, I believe the adoption of ‘social kinds’ language in ontology is deceptive since there is no agreement on the metaphysical basis of social kinds or how the social relates to the natural. The best theoretical models available need to guide classification so that the category of race picks out some ontological reality (i.e., a ‘real kind’), but they are too weak to provide satisfactory answers. The metaphysics of race lacks a firm basis for understanding where such demographic ‘joints’ are located, and without such a firm basis the debate could only be about conceptual schemes—a debate over systems of racial categorization.

Chapter 4: What is Race?

The entirety of the metaphysics of race debate could be seen as a response to the question: *what is race, really*? Today, race is still a significant dimension of social life, at least in many societies, and continues to organize populations. That much seems indisputable. What remains open to question is how race should be understood. Based on my discussions in the previous chapter, I raise the following investigative questions to differentiate the positions I focus on in the metaphysics of race debate:

1. Are races sets (arbitrarily constituted) or classes (conceptually constituted)?
2. If races are classes, are they conventional categories (concepts pertaining to language and theory) or real kinds (entities pertaining to the world)?
3. If races are real kinds, are they merely natural kinds (grounded in the structure of natural reality) or social kinds (grounded in the structure of social reality)?
4. If races are social kinds, are they grounded in agency (rule-based) or grounded in the practical circumstances of social agents (based on tacit dispositional strategies)?
5. If races are rule-based social kinds, are they based on explicit rules or implicit rules?

These are but a selection of questions to raise, but their combined answers should express the underlying commitments of distinct positions in the metaphysics of race.

4.1: Camps of the Debate

The metaphysics of race contains a variety of divergent ontological positions and it would be impossible to do its diversity justice in my exposition. Here, I limit myself to considering *What is Race? Four Philosophical Views* (2019) by Joshua Glasgow, Sally

Haslanger, Chike Jeffers and Quayshawn Spencer. An honorary addition to this company is Michael Hardimon, whose work is in dialogue with these philosophers. Since the entirety of the debate is beyond the breadth of this thesis, I present only a narrow expanse of the arguments corresponding to the abovementioned investigative questions in giving context to the study of related thought experiments.

4.1.1: Realism

Racial realism is the camp that argues races are ontologically real. The first school of realism I consider is naturalism. It includes Quayshawn Spencer's non-essential naturalism, as well as Michael Hardimon's minimalism and populationism. The second school of realism I consider is social constructionism. It includes Sally Haslanger's socio-political constructionism and Chike Jeffers' cultural constructionism.

4.1.1.1: Non-Essential Naturalism. Quayshawn Spencer argues for a pluralist, non-essential naturalist account of race, and claims that 'race' need not be racist. There are different ways of understanding race, some of which are social or biological. Indeed, he does not argue against social scientific theories of race. What he wants to show is that metaphysicians cannot outright declare that something like 'race' does not exist based on their (mis)understanding of biology or sciences in general. Anything in a successful scientific theory that matches the common use of 'race' well enough to be useful is going to be what 'race' is in that theory. He does not deny the possibility that a theoretical conception could be ethically problematic, but he does emphasize how scientific misappropriation of 'race' has always been discredited by science itself.

In terms of my investigative questions, Spencer argues that (1) races are classes (conceptually constituted); (2) races are conventional categories (concepts pertaining to language and theory) that ground real kinds (entities pertaining to the world); (3) and races are natural kinds (grounded in the structure of natural reality).

(1) Races are classes (conceptually constituted) because they share biological features that are consistent with ancestry groups. Ruling out all other conceptions, in population genetics they are epistemically useful and conceptually justified (Spencer, 2019a, p. 77). In the United States race is ordinarily understood in terms of OMB [the Office of Management and Budget] official scheme, classified according to presumed major continental ancestry groups (pp. 78, 92, 94). Incidentally, the OMB races align enough with the continental groups of population genetics to allow self-reporting to be an accurate biological predictor for researchers (pp. 100-103). Spencer calls this useful alignment "the identity thesis" for tracking the same "genomic ancestry" in both cases (pp. 100-103).

(2) Races are conventional categories (concepts pertaining to language and theory) because they are classes theoretically generated to capture real kinds in nature (entities pertaining to the world). For Spencer, when categories belong to an empirically successful scientific theory, they have reality. Since race, as a category, "adequately captures the collection of entities that are actually used in empirically successful biology" it has biological reality (Spencer, 2019a, p. 77). Spencer describes his position as "radically pluralist" and argues that "there isn't [only one] dominantly correct answer to the question of what race is and whether it's real in the relevant context, but there's still at least one dominantly correct answer to this question" (pp. 211, 213). In other words, there are different

competing accounts of what a race is, and successful science helps to show which account should be taken more seriously.

(3) Races, as real, are natural kinds (grounded in the structure of reality) as categories of natural entities in population genetics. Spencer is a trained scientist and does not overlook the questionable use of 'natural kinds' outside theoretically justifiable contexts. He limits his use of the HPC theory only to his analysis of the Noah Rosenberg et al. (2002) study of the 'K=5' genetic clusters account of race (pp. 96-103). According to Spencer, the ontological debate over race is not a debate over its status as a natural kind, but a linguistic debate about whether the "American English speakers' ideas about race form a natural kind" which is "highly suspect and almost certainly false" (p. 236). Something is a natural kind only if it is captured by a projectable predicate used for the inductive inferences of an empirically successful scientific theory (and everything else, including my fourth and fifth investigative questions, are a distraction).

4.1.1.2: Minimalism and Populationism. Michael Hardimon argues for a minimalist and a populationist account of race and claims that 'race' need not be racist. To be clear, he is not a pluralist about race. There are pernicious conceptions of race that he calls "racialist" which should not be confused with the concept of race itself. According to Hardimon, the ordinary concept of race has a "logical core" called the "minimalist concept" that accurately captures its intension and extension (2017, p. 57). This "logical core" only captures what is "rational" and "nonracialist" about "the ordinary concept, which does not invoke the idea of intrinsic biological essences or normatively important features, nor does it posit a correlation between such features and visible physical characters (p. 3). It maintains that races are distinguished by differences in patterns of “visible physical features” that correspond to “differences in geographic ancestry” (p. 3). In other words, certain features indicate ancestry, which in turn signals a relation to ancestral geography.

Populationist race adds scientific concepts (of genetic transmission, phenotype, reproductive isolation, and founding population) to minimalist race for tracing back geographically separated and extrinsically reproductively isolated founding populations in genetics (Hardimon, 2017, pp. 3, 99). In other words, it is the concept used to study the biological evidence for differences in geographic ancestries assumed in the minimalist concept.

Hardimon also adds “the concept of socialrace” to the mix, which is his name for the reappropriation of the “racialist concept of race” as a “nonracialist, critical, emancipatory concept of social groups” (p. 3). The racialist concept “maintains that races have intrinsic biological essences, are distinguished by normatively important features such as intelligence and moral character, and can, based on these features, be objectively ranked as superior and inferior” (p. 2). The concept of socialrace preserves from this racialist concept the reference to the social position occupied by particular social groups, or the system of social position, for emancipatory purposes (p. 131). Hardimon’s socialrace is virtually the same as Sally Haslanger’s socio-political concept of race, which I discuss later.

In terms of my investigative questions, Hardimon argues that (1) races are classes (conceptually constituted); (2) races are real kinds (entities pertaining to the world); (3) races are natural kinds (grounded in the structure of natural reality) and social kinds (grounded in the structure of social reality); (4) and as social kinds, races are grounded in practical circumstances (based on tacit dispositional strategies).

(1) Races are classes (conceptually constituted) because they share phenotypic features in common as much as they share their geographic ancestry. The minimalist core of the ordinary concept of race classifies groups of human beings according to the way differences in their geographical ancestry are signalled by distinguishable patterns of visible physical features (Hardimon, 2017, pp. 3, 31, 37). Further, "populationist race" is the "scientization" of this minimalist core concept with the addition of genetic criteria for the biological inheritance of variation from historically distinct and reproductively isolated founding populations (pp. 3, 99, 113). Hardimon also introduces the concept of "socialrace" which is a social group that shares normatively significant features, unrelated to the minimalist concept (p. 131).

(2) Races are real kinds (entities pertaining to the world) because they are categorizing real classifications found in the human population. The minimalist concept of race is "grounded on a few biological differences of a comparatively superficial nature" (i.e., colour and physiognomy corresponding to geographical ancestry) (Hardimon, 2017, pp. 37, 78). In contrast, the reality of the populationist conception depends on the ancestral lines of descent studied by population genetics (p. 120). It is important to note that Hardimon takes this populationist concept to be theoretically identical to the minimalist race in science, claiming that "minimalist race = populationist race is analogous to the claim that water = H2O" because minimalist race and populationist race are "alternative conceptual

representations of one and the same kind" (p. 120). This relation of theoretical identity makes it clear that Hardimon subscribes to the Kripke-Putnam view of natural kinds, as having their theoretical essences fixed via causal and social practices. Additionally, Hardimon argues that "socialrace" has social, rather than biological, reality (p. 131). So, socialrace, although real, is not the representation of the same kind as the minimalist or populationist race.

(3) Races, as minimalist and populationist, are natural kinds (grounded in the structure of natural reality) because they are categories of real entities, while "socialraces" are social kinds (grounded in the structure of social reality) since they merely categorize social distinctions. I must say that the relationship between natural kinds and social kinds for Hardimon does not make a difference here. Minimalist race is grounded in distinguishable patterns of visible physical features corresponding to differences in geographical ancestry (Hardimon, 2017, p. 37). Populationist race is grounded in the relation of these geographical patterns to population genetics (pp. 58-62). On the other hand, "socialrace" is grounded in the racialized patterns of social position found in social hierarchies of power (pp. 131-135).

(4) As a social kind, "socialrace" is grounded in practical circumstances of racial ideology (based on tacit dispositional strategies of agents). Socialraces are social groups with differential positions of power that are defined by normative value assignments to their normatively neutral physical properties (Hardimon, 2017, p. 131).

4.1.1.3: Socio-Political Constructionism. Sally Haslanger argues for a socio-political constructionist account of race, and that 'race' is mostly racist. To be clear, she is a pluralist, but she does not avowedly disclose the extent of her pluralism about race. She accepts that race could be understood through a different lens, but that it should primarily be understood in terms of a social caste/class system that privileges some while subordinating others within an unjust social hierarchy.

Haslanger's project is purportedly "ameliorative" because it seeks to identify what legitimate purposes some might have (if any) in categorizing people by race, and it seeks to develop conceptions for the achievement of emancipatory ends (2012b, p. 366). She is not interested in the best description of race in any or all cases. The details of her project are not important here. What is relevant is that she is, instead, interested in revising conceptions of race for situated political purposes.

It is important to make a note of Haslanger's pragmatic/pluralist approach to the concept of race. In response to Quayshawn Spencer's argument, Haslanger claims to be "happy to grant that Spencer has pointed to one way of understanding race in response to one set of questions and concerns" and concedes that if there is even "a chance" of finding a relevant biological difference, "we should be doing research that explores such a possibility" (2019a, pp. 153, 156). She makes this clear in the following:

> It is compatible with my view that there are genetic ancestry groups corresponding to the OMB categories and that we should keep an open mind about whether there are significant medical results to be had by doing genetic research on these groups. Given the challenges of gathering data and the significance of racial identities in the United States, it may also be fruitful to

> continue to use the term ‘race’ for these categories. Such questions and concerns are very different from my own, however, so if concession on these points is what [Spencer] is after, I see no conflict with my view (with the substantial caution that it is an empirical question whether using the term ‘race’ for these categories is, on balance, politically wise in the long term, or not). (Haslanger, Haslanger’s Reply to Glasgow, Jeffers, and Spencer, 2019a, pp. 153-154)

This should not be mistaken as a purely biological position but as an acknowledgement of her radical pragmatism/pluralism about race. The socio-political practices that Haslanger wants “to describe and explain concern patterns in human interaction, racial ideology, and durable forms of social stratification” alone, not the patterns in nature that Hardimon or Spencer concern themselves with (Haslanger, 2019a, p. 156). She is not interested in those projects. There are different conceptions of race and both the socio-political and biological conceptions belong to very different contexts of inquiry and are made for very different theoretical purposes (p. 156). The view is that there are physical differences amongst humans associated with race, but race itself is when these differences justify systematic power relations. So, Haslanger acknowledges race as a natural kind, yet, she only indulges the debate about race as a social kind.

In terms of my investigative questions, Haslanger argues that (1) races are classes (conceptually constituted); (2) races are both conventional categories (concepts pertaining to language and theory) and real kinds (entities pertaining to the world); (3) races are social kinds (grounded in the structure of social reality) or, possibly, natural kinds (grounded in the structure of natural reality); (4) and as social kinds, races are grounded in agency (rule-based); (5) and, as rule-based social kinds, races are determined by implicit rules.

(1) Races are classes (conceptually constituted) because they share in common several epistemically and pragmatically relevant standards (Haslanger, 2019b, pp. 16-18, 33). Here, I mention what I believe to be the two most important theoretical distinctions made by Haslanger. Firstly, the concept of race that users typically take themselves to be applying is called "the manifest concept" of race, while the concept of race determining how users apply the term to different cases is called "the operative concept" of race (1995, p. 102). Often enough, the manifest concept and the operative concepts drift apart, which could make it seem as if race is arbitrarily constituted. However, since racial practices remain regular, the operative concept continues unchanged. Secondly, Haslanger advocates for the use of the term 'colour' to refer to the physical markers of human bodies, and she advocates for the use of the term 'race' to refer to conferred normative positions of 'colour' in racial practices. In parallel to the feminist slogan "gender is the social meaning of sex," Haslanger claims that "race is the social meaning of the 'coloured,' that is, geographically marked, body" (2012a, p. 308). In this sense, 'colour' classification is based on contextually variable associations of observed or imagined physical markers to geographic ancestry (2019b, p. 25). This includes "any cluster of physical traits that are assumed to be inherited from those who occupy a specific geographical region or regions" such as "eye, nose, and lip shape, hair texture, physique, and so on" (2012a, p. 307). In contrast, race merely classifies 'colour' markers in a socio-political order, without directly referencing ancestry. It is only within the context of a racial ideology that 'colour' becomes a mark for differential treatment and

social power (subordination/privilege) (2019b, pp. 25-26). Therefore, 'race' and colour' are respective analogs of social and physical realities at play, even though the latter is no more informative than to associate bodies with ancestry. I must highlight, as Haslanger seems to be aware, that the analogy between 'colour' and 'race' leaves open the concern for the former to become a similar socio-political issue.

(2) Races are both conventional categories (concepts pertaining to language and theory) for being epistemically relevant distinctions in the social sciences and real kinds (entities pertaining to the world) for categorizing real positions within the social hierarchy. Racial categories are flexible enough to meet different needs in different social contexts (Haslanger, 2019b, pp. 21-22, 33-34). As real kinds, races are differential positions of power within a social hierarchy, and "we discover these kinds through empirical inquiry, just as we discover chemical kinds through empirical inquiry" (pp. 5, 25-26). Such discovery of hidden natures makes it clear that Haslanger subscribes to the Kripke-Putnam view.

(3) Races are social kinds (grounded in the structure of social reality) because they depend on social practices (Haslanger, 2019b, pp. 25-26). According to Haslanger, social kinds are "kinds of things that exist in the social world (and so, in some sense, depend on us)" (p. 5). In her Kripke-Putnam view, the referent of 'race' is fixed by socio-linguistic practices organized by a linguistic division of labour. However, the degree to which her analysis goes beyond mere linguistic practices in society remains unclear. Also, it must not be overlooked that, for Haslanger, races remain possible natural kinds (grounded in the structure of natural reality), even though she does not engage with them as such.

(4) As social kinds, races are grounded in agency (rule-based) because they operate by the rules outlined by Haslanger's socio-political account of race:

> [A] group G is racialized relative to context C iff members of G are (all and only) those (i) who are observed or imagined to have certain bodily features presumed in C to be evidence of ancestral links to a certain geographical region (or regions)—call this "colour"; (ii) whose having (or being imagined to have) these features marks them within the context of the background ideology in C as appropriately occupying certain kinds of social position that are in fact either subordinate or privileged (and so motivates and justifies their occupying such a position); and (iii) whose satisfying (i) and (ii) plays (or would play) a role in their systematic subordination or privilege in C, that is, who are along some dimension systematically subordinated or privileged when in C, and satisfying (i) and (ii) plays (or would play) a role in that dimension of privilege or subordination. (Haslanger, Tracing the Sociopolitical Reality of Race, 2019b, pp. 25-26)

(5) As rule-based social kinds, races are determined by implicit rules because they depend on the operative concepts that are at play in racial practices without having to be transparent to agents. For Haslanger, the socio-political account outlined above highlights key aspects of racializing practices that may be "occluded or masked" and enacted "mindlessly" (2019b, pp. 33-34).

4.1.1.4: Cultural Constructionism. Chike Jeffers argues for a monist cultural constructionist account of race, and that 'race' need not be racist. He believes that race is a cultural mindset, experienced from the inside by inhabiting a unique worldview, in sharing a way of life and cultural practices informing the historical experiences of its members.

Jeffers, much like Haslanger, makes a concession to Spencer in admitting that "one cannot tell the story of racial distinctions without biological diversity entering the picture"; and that the "forms of physical difference involved in racial distinctions are necessarily at least partially related to forms of reproductive isolation, whether as a result of people being geographically separated going back to the distant past or through more recent social distinctions"; and that "we" must be as "open-minded [as Spencer] about what such research might uncover regarding medical matters and how we might connect what is uncovered to our usage of common sense racial categories in medical settings" (Jeffers, 2019b, p. 182). But unlike Haslanger, Jeffers argues that "race is fundamentally social and not fundamentally biological" and that "biological continuities and discontinuities between human populations is a subject independent of the social recognition of racial differences" (pp. 181-182). In other words, the biological reality of race does not run parallel to, nor is it the cause of, the social reality of race. Instead, it is itself the evidence for historically divergent socio-cultural practices.

In terms of my investigative questions, Jeffers argues that (1) races are classes (conceptually constituted); (2) races are conventional categories (concepts pertaining to language and theory); (3) races are like social kinds (grounded in the structure of social reality) but also like real kinds (entities pertaining to the world); (4) as likened to social kinds, races are grounded in agency (rule-based); (5) and as rule-based social kinds, races are determined by explicit rules.

(1) Races are classes (conceptually constituted) because they share unique “traditions” and “ideals of life” in common (Jeffers, 2019a, pp. 49-50). They are “appearance-based groups that initially result from the history of Europe’s imperial encounters” (p. 65). But, since then, their shared culture continues to organize them in close proximity to the original group of peoples living in unequal power under the structures imposed by the European age of empires (p. 62).

(2) Races are conventional categories (concepts pertaining to language and theory) because they shape the minds/values of their members, but real kinds (entities pertaining to the world) as civilizational blocks in history. Once European imperialism divided people into groups of unequal power, racial categories shaped the identities of different racial groups by making their members view those very racial categories as culturally significant over time (Jeffers, 2019a, p. 62). But races are also real kinds because such “social distinctions lead to inhabiting relatively different worlds and thus participating in different ways of life” (pp. 50, 63).

(3) Races are like social kinds (grounded in the structure of social reality) because Jeffers’ cultural constructionism views shared “traditions” and “ideals of life” as the most important defining factors of races (2019a, pp. 49-50). These factors are so important that they explain breeding patterns which have effectively shaped and continue to shape race as a natural kind (grounded in the structure of natural reality). It is not made clear whether Jeffers subscribes to the HPC or the Kripke-Putnam theory of kinds, even though his view is at least somewhat aligned with the HPC account.

(4) As social kinds, races are grounded in agency because their members need to be active in their cultural practices to a minimal degree. This is because, for many of their members, identification with the racial group is connected with their investment and engagement in practices that they take to be distinctively related to the group's existence (Jeffers, 2019a, p. 66).

(5) As rule-based social kinds, races are organized explicitly by cultural distinction. According to Jeffers, three key cultural commonalities organize the psychology of race: the cultural experience of racial consciousness itself; racial consciousness as facilitating new cultural developments; and racial consciousness as shaped by prior cultural developments (2019a, pp. 64-66). As such, shared cultural practices and racial solidarity are embedded in the historical contingencies of racial (self) consciousness, transparent to racially conscious members (p. 62).

4.1.2: Anti-Realism

Racial Anti-Realism is the theoretical camp that argues races are not ontologically real. I will only consider one philosophical position in this camp. Joshua Glasgow argues for an eliminative position that is reminiscent of racial skepticism but with the novel addition of a disjunction that Glasgow calls "basic racial realism."

4.1.2.1: Eliminativism. Joshua Glasgow argues for an eliminativist account of race, and that 'race' could avoid being racist. He is not a pluralist, but he pushes for an elaborate disjunctive position to demotivate racial realists—such that "ultimately we'll have to choose between the idea that race is an illusion and the idea that race is real in a basic, scientifically irrelevant sense" (2019b, p. 118). He argues that race does not exist as a biological nor a social kind (p. 117). Glasgow understands 'race' to be an unrepairable mistake of ordinary thinking about the visible physical features of humans (p. 262). He argues for an unanalyzable kind of racial Realism that only rests on distinctive sets of “visible traits” (p. 246). This he calls “basic racial realism” which holds “that race is real in a way that is more ‘basic’ than what science aspires to” (p. 139). As strange as it sounds, Glasgow claims that race is a “basic kind” of real existents like “stuff around trees” (p. 139). Additionally, in line with Spencer, he argues that his basic kind conception of race could serve “as an imperfect proxy” for some scientifically relevant kinds (p. 141). The only difference between this proxy relation and the one advanced by Spencer is that Glasgow does not take this to be an identity relation, but one of happenstance.

In terms of my investigative questions, Glasgow argues that (1) races are either sets (arbitrarily constituted) or, possibly, classes (conceptually constituted); (2) if races are classes, then they are, possibly, real kinds (entities pertaining to the world); (3) if races are real kinds, then they are, possibly, natural kinds (grounded in the structure of natural reality).

(1) Races are either sets (arbitrarily constituted) erroneously thought to be in reality, or possibly classes (conceptually constituted) of perceptible similarities between bodies. As sets, races are defined by a list of conjunctive visible traits presumed to be inalterable and inheritable, regardless of the similarities and differences that run across different sets (Glasgow, 2019b, pp. 124-125). As classes, races share common perceptible features that may be explicated by studying the use of the term in ordinary linguistic practices (pp. 115-116).

(2) Races are possibly real kinds (entities pertaining to the world) but only as what Glasgow calls unanalyzable “basic” kinds (2019b, pp. 118, 139). In other words, they exist as stable patterns of error in human judgement.

(3) Races are possibly natural kinds (grounded in the structure of natural reality) but, oddly enough, not in a biological or social sense. They are merely unified as perceptible similarities and still ‘natural’ (perhaps) due to the ‘nature’ of perception. However, they are scientifically useless (unless in the science of perception) and, for this reason, Glasgow calls them “basic kinds” (2019b, p. 139). That is to say, there are “basic kind” existents in reality, like “stuff around trees” (p. 139). Since “stuff around trees” could be, according to Glasgow, considered to be a basic kind, so should race (p. 143).

4.2: A Critical Overview

The overview of the abovementioned arguments displays a striking pattern. With the help of the investigative questions, I have outlined the metaphysical peculiarities of each position in ‘Table 1,’ below. Perhaps the most important point to note is how Haslanger, Glasgow and Jeffers excuse themselves for silently accommodating Spencer’s biological argument (I exclude Hardimon, momentarily, since his position unreservedly accommodates Spencer’s view). Jeffers’ position accommodates the kind of biological difference between races that Spencer argues for, but he views such biological differences as evidence for divergent cultural practices which have historically given rise to the contrast between the races. This is an interesting and persuasive hypothesis that attempts to explain biological differences of race as the by-product of different cultural practices, particularly those influencing breeding habits. However, the same cannot be said about the concessions made by Haslanger and Glasgow. Haslanger, more or less, concedes to Spencer's argument, conditioned on the same empirical and epistemic standards that Spencer holds for himself. She takes him to be right, even though she is only interested in the socio-political aspect of race. On the other hand, Glasgow reiterates Spencer's argument short of the identity thesis, with the 'basic kind' notion of race instead of the OMB scheme. Yet, he fails to explain why a 'basic' kind has no scientific use when he finds it appropriate to use it as a proxy for geographic ancestry in population genetics.

Table 1:

The metaphysics of race debate according to the investigative questions in Chapter Four

	Quayshawn Spencer	Michael Hardimon	Sally Haslanger	Chike Jeffers	Joshua Glasgow
School of Thought	Metaphysical Realism	Metaphysical Realism	Metaphysical Realism	Metaphysical Realism	Metaphysical Anti-Realism
Theory	Non-Essential Naturalism	Minimalism and Populationism	Socio-Political Constructionism	Cultural Constructionism	Eliminativism
Sets					✓
Classes	✓	✓	✓	✓	?
Categories	✓		✓	✓	
Real Kinds	✓	✓	✓		?
Natural	✓	✓	?	✓	?
Social		✓	✓	✓	
Rule-based			✓	✓	
Dispositional	✓				
Explicit				✓	
Implicit			✓		

In their response to Spencer, Haslanger and Glasgow seem to undermine the interpretive capacity and explanatory power of their arguments. Importantly, Haslanger does not explain what metaphysical difference it would make to her position if Spencer is wrong. The concern is that if it makes no difference, their arguments have nothing to do with one another. On the other hand, Glasgow needs to explain why scientists could possibly find his ‘basic’ notion a useful proxy in their research method.

I like to think that Glasgow’s ‘basic kind’ is an inside joke, one testing the metaphysical competence of philosophers. The doctrine of kinds is highly controversial and I do not advocate it. All the same, I am not blinded to Glasgow’s attempt to repackage the colloquial use of ‘kinds’ as yet another technical variant of real kinds in metaphysics. Real kinds/natural kinds/social kinds are not simply classes of existents like “stuff around trees” unified by the mere similarity of being around trees. Metaphysical kinds are unique categories of real entity types with modal implications for their members. If ‘stuff around trees’ is a basic kind, then basic kinds are not unique categories, but only sets of objects. But even if they were unique for being basic categories, they carry no modal implications for their members. Consider ordinary salt as a chemical kind, every instance of which is necessarily going to be salty and chemically made of the compound Sodium Chloride. Compare it to stuff around trees as a basic kind. Do samples of stuff around trees share invariant empirical properties? Instances of stuff around trees are not inductively useful, because an undefined perimeter ‘around’ trees could extend over everything in the universe (i.e., all beings are stuff around trees). But I suspect that Glasgow knows this already. If ‘stuff around trees’ has no modal implication nor is it inductively projectible in hypothesis, it is rebranding colloquial ‘kinds’ to prove a point: there is confusion about what metaphysical ‘kinds’ are meant to be in the metaphysics of race. Besides, if metaphysicians were to fall for the notion of basic kinds, they would have a much more parsimonious ontology of race that does not rely on biological or sociological posits. By claiming that race is an unanalyzable pattern of error in perceptual judgement, Glasgow satisfies the analytic metaphysician's aesthetic "taste for desert landscapes" (Quine, 1948, p. 23). However, if the 'basic kind' option does not sit too well with the connoisseurs of minimalism, eliminativism presents an even safer option advocated by Glasgow. A clever argument, but one relying on the failure to ask what a 'kind' is meant to be in metaphysics.

In a broader sense, there seems to be something odd about the entire debate. In one respect, different camps seem to be talking past each other, while in another respect they seem to be in harmony. Regarding her interlocuters (Spencer, Jeffers and Glasgow), Haslanger states that "we disagree not only in our conclusions, but also in our methods [and we] are engaged in different projects that overlap, but shouldn't be expected to yield the same results" (2019b, p. 152). This is an outstanding admission, for raising the possibility (and I suspect in some ways the actuality) that in this debate there is little dialogue taking place over ontological issues. As she duly notes, Glasgow and Spencer "focus primarily on race as a classification of humans" regardless of social and cultural practices, while Jeffers and Haslanger mainly "focus on a broader range of social and cultural practices" without addressing the competing classifications of race (p. 152). For Glasgow and Spencer, there are right and wrong ways of understanding racial classification, and racial relations are of secondary concern. For Jeffers and Haslanger, there are right and wrong ways of understanding racial relations, and racial classifications are of secondary concern. These are completely different theoretical projects, even if tightly intertwined. Haslanger notes that these different starting points are pushed further apart in their method. This is why, overall, Glasgow and Spencer examine whether racial classifications fit the world, while Jeffers and Haslanger examine why certain social and cultural practices are thought to be racial (p. 152). Divergent questions and methods leave only the theme and terminology in common.

From a different point of view, it is easy to see how these positions are in harmony since they all map onto the account of race outlined by Hardimon. Glasgow is concerned with what Hardimon calls "minimalist race" (i.e., race as a basic kind); Spencer is concerned with "minimalist race" (i.e., race in terms of the OMB scheme) matching "populationist race" (i.e., race as a classification in population genetics); and Haslanger is concerned with "socialrace" (i.e., race as socio-political). The only outlier is Jeffers who, nonetheless, acknowledges the compatibility of his work with "socialrace" much like Haslanger, but without her reliance on institutions and her reservations for an independent ontology of the biological determinates of race. Still, Jeffers does commit himself to Hardimon's "logical core" of the concept of race to bridge the gap between "visible physical features" and ancestry in ordinary common sense ideas and lived experiences of race (2019a, pp. 39, 43). Given that Hardimon's theory articulates these series of distinct but interrelated race concepts to "provide a language that makes it possible to think and speak coherently about race," and given how the positions of the theorists considered more or less fit into distinct compartments of Hardimon's theory, two observations can be made. Firstly, since Glasgow, Haslanger and Jeffers have made concessions to Spencer's naturalistic argument in matching the ordinary concept of race to its biological proxy, they have also made similar concessions to Hardimon's parallel link between the "minimalist" and the "populationist" concepts of race. Secondly, since Spencer does not deny the social dimension of race, he has to accept something like Hardimon's "socialrace" concept. Subsequently, if Glasgow, Haslanger and Jeffers accept something like the “minimalist” and the “populationist” concepts, in addition to accepting “socialrace,” and Spencer accepts “socialrace” in addition to something like the “minimalist” and the “populationist” concepts, then they would all be in harmony under Hardimon’s theory. It is as if they all accept Hardimon’s conceptual distinctions, but disagree about how much they do so, the method they prefer to adopt, and what aspect of the theory should take priority in thinking about race. So, importantly, they hold compatible views but do not say so outright.

Chapter 5: The Theory of Make-Believe

This chapter introduces the theory of make-believe in some detail. My focus will be on providing an abridged exposition to help make sense of the main argument in my thesis. However, Walton uses familiar terms in quite distinctive ways, and clarifying what exactly he means by certain key terms of art is going to need additional explanation. Understanding Walton's theory can be advantageously supplemented by a few background connections. So, I begin with the theoretical orientation and follow through with a working guide for the technical details. Still, I should make clear from the outset that I do not offer an overall application of Walton's theoretical toolset. Rather, I adopt and adapt the most relevant aspects of his framework to develop a powerful methodology for thought experiments.

5.1: Thought Experiments as Make-Believe

In his magnum opus, *Mimesis as Make-Believe: On the Foundations of the Representational Arts* (1990), Kendall Walton offers the most comprehensive account of his theory. But what is his theory called? Walton does not give it a clear label. An interesting name has been introduced by Stacie Friend, in "Imagining Fact and Fiction" (2008). Friend coins 'Walt-fiction' to label Walton's unique way of delineating 'fiction' in contrast to its broader meaning because, as she observes, Walton is only interested in the works of fiction that prompt imaginings in games of make-believe (pp. 153-154). 'Walt-fiction' has gained some currency, but I believe Friend misses the mark in two ways: firstly, Walton's is not merely a theory of fiction, but also a theory of the representational arts; and secondly, 'Walt-fiction' suggests that there is something unusual about the subject of Walton's theory, but there is not. Walton is interested in the very same things that usually count as art and fiction. The eccentricity of his thought comes not from the target of his theory but from the novelty of his approach. At least in my view, the best description of Walton's theory comes right out of his book title, which is why I refer to it as the 'theory of make-believe.'

Walton's inspiration comes from the way children engage their imagination when they play with toys. He calls this mentally engaged activity a 'game of make-believe' and makes a strong case that the adult activities in which art and fiction are embedded and made meaningful are best seen as continuous with the games children play with toys (Walton, 1990, p. 11). This theoretical move collapses the distinction between the so-called 'representational arts' (e.g., statues, paintings, photographs) and 'works of fiction' (e.g., novels, plays, poems) to argue that any **representational object** *functions as a prop to equip imaginings* (p. 11). Games of make-believe are social, and representational objects have the social function of facilitating games of make-believe (p. 69). More accurately, he proposes that the two share the same function, and this function defines what a representational object is.

Beyond the representational arts and literary fiction, Walton's theory of make-believe presents a handy framework for understanding thought experiments. In this theory, disagreements over the use of thought experiments are similar to disagreements over art and literary criticism. Thought experiments are represented as fictional narratives, but the sort of fiction that prioritizes an epistemic purpose (Meynell, 2014, p. 4150). Walton's conception of fictionality tracks epistemic value which makes it perfectly suited for thought experiments. A thought experiment may begin with some phrases like "Imagine that…" or "Suppose it that…" as a way of mandating participants to imagine what follows. For Walton, what follows (e.g., a proposition) is fictional in so far as it is *a 'fictional truth*,' and **fictionality** is *the quality of being a 'fictional truth*' (1990, p. 35). In turn, **fictional truth** is simply whatever *prescribes, mandates or prompts imaginings in the context of make-believe* (p. 39). If I were to imagine the wrong thing, my wrong imagining would not be a fictional truth. It could be said that 'fictional propositions' are those propositions that are to be imagined—whether or not they are in fact imagined (p. 39). However, the talk about 'fictional propositions' makes it sound as though they are propositions (p. 41). But they are only a linguistic/logical shorthand[25] for clarifying what is to be imagined in some fiction (p. 36). For example, about a fictional work stating that X is the case, one can say "it is fictional that X" and "it is not fictional that ~X" in propositional form to clarify what is fictionally true in a game of make-believe (p. 35). This helps participants to enumerate what has fictionally happened or is fictionally happening in case of any confusion (Meynell, 2014, p. 4158). So, to harness the epistemic value of thought experiments, participants are to respond as if the make-believe was really true.

Fiction and *truth* are not opposites for Walton. They are *analogs*: **fictionality** is to imagining how *truth* is to *belief* (Walton, 1990, p. 41). In other words, **being fictional** *is analogous to being true*, while *imagining something is analogous to believing something*.

[25] The theory of make-believe is not, technically speaking, a linguistic approach. In an earlier phase, Walton did concern himself with 'fictional truth value' and 'fictive sentential operators' (1973, pp. 287-288). But, in the theory of make-believe, careful linguistic analysis becomes a way "to inform" rather than making "propositions fictional" (1990, p. 130). He states: "Readers who reject propositions or prefer to understand them differently are invited to reformulate my claims about fictionality however their philosophical conscience dictates—in accordance with their preferred way of treating (so-called) propositional attitudes generally. It is my belief that any reasonable reformulation will be recognizably the same, that the substance of the problems it treats and its ways of treating them will remain" (p. 36). Unfortunately, the talk of truths and propositions could easily lead to the mistake of incorporating their associated theoretical baggage into the theory.

Likewise, fictional propositions should not be thought of as non-factual. For example, it is true in the fictional world of *War and Peace* that Napoleon invaded Russia, while it is also actually true that, in the real world, Napoleon did invade Russia (p. 79). Every fictional representation, including thought experiments, will contain fictional truths that are true of the actual world (Meynell, 2014, p. 4159). It is important, therefore, to emphasize how **fictional truths** *are not like truths* for Walton, regardless of their verisimilitude, but mandate one to imagine something (1990, pp. 41-42). They *determine fictional content*, whether or not they match what is true in the actual world. But *if they match actual truths, then they are also true of the actual world*.

In Walton's theory, '**fictional worlds**' *are not 'possible worlds' of metaphysics* (1990, pp. 64-67). Unlike possible worlds, conceived as totalities of determinable propositions of maximally consistent sets of facts, most fictional worlds are open and not fully determinate (Meynell, 2014, p. 4160). Works of fiction only provide enough of their fictional world to work with. Fictional worlds *are composed of work worlds and game worlds*. *What the representational object (i.e., the prop) provides* for generating make-belief is called the '**work world**,' which includes the fictional narrative of the thought experiment, and *what the participant provides* for generating the fictional world over and above the work world is called the '**game world**' (p. 4159). In short, the fictional work is an object in the actual world, and its fictional world is generated by participants within a game of make-believe, with no additional metaphysical baggage required.

The theory of make-believe is metaphysically deflationary because it views the imaginings of participants to generate make-believe. The first step of the process involves a narrative prescribing a fictional scenario to be imagined. Then, in response to the narrative, the participant generates the prescribed imaginings and, in so doing, enables the thought experiment's scenario in make-believe. There is nothing more and nothing less metaphysical here, besides the engagement of participants' imagination. Of course, both the narrative and the participant could fail to meet expectations. Hence, any imagined results yielded from thought experiments ought to be deemed tentatively true, if true at all.

5.2: Mimesis as Make-Believe

One helpful way of being introduced to Walton's theory is by becoming familiar with a thesis advanced by Ernst Gombrich, which Walton discusses in "Pictures and Make-Believe" (1973) and "Pictures and Hobby Horses: Make-Believe Beyond Childhood" (2008). In "Meditations on a Hobby Horse or the Roots of Artistic Form" ([1963] 1978), Gombrich uses the example of an ordinary toy—the humble hobby horse—to argue that the notion of 'representation' in the fine arts is best understood as a synonym for 'substitution' and the verb 'to represent' is best understood as 'to substitute for' (p. 1). In other words, representational objects are said to represent not by merely 'standing in' for something, but by playing a role in a certain context. Here, the idea[26] is that a representational object, whether it be a hobby horse, the painting of a landscape or a fictional narrative about war, functions adequately enough as a substitute in some way, whether it be a horse to ride on, a landscape to admire or a war to endure. As such, representational objects substitute the need for something by fulfilling enough of its role in a game of make-believe.

[26] A helpful analogy to consider is how a substitute teacher becomes the representative of a class of students (e.g., so and so's class) by replacing that regular teacher.

I call Gombrich’s account a substitution model of representation. In the right context, a representational object plays the role that is needed or demanded of it, just as counterfeit coins and makeshift keys can play the role of real coins and authentic keys to get the job done (Gombrich, [1963] 1978, p. 4). Relying on the same substitution model, Walton names any object that functions in a substitutive capacity a ‘prop’ simpliciter (1990, pp. 11, 35-43, 96). For him, a representational object is a **prop** because it *plays the specific role demanded of it in a game of make-believe*.

The substitution model is embedded in Gombrich’s theory of mimesis. Usually translated from Greek as ‘imitation,’ mimēsis has deep philosophical roots (Preus, 2007, p. 173). The notion has traditionally been understood to mean ‘*re*-presenting’ reality, which underpins the aesthetic theory of Representationalism (Wolterstorff, 2015, p. 671). In time, the rise of Romanticism in Western art gave way to the view that saw mimesis as the expression of genius from the artist (p. 671). Gombrich offers (or revives) an alternative: while Representationalism views mimesis as the *mirroring* of *external* reality, and Romanticism views mimesis as the *expressing* of *volitional* reality, for Gombrich mimesis is the *conjuring* of a *substituted* reality. In other words, representational objects are neither copies (Representationalism), nor revelations (Romanticism). They are like replacements or surrogates in some mimetic arrangement that fills a niche. In correspondence, Walton calls this arrangement a ‘game of make-believe.’ Just as a substitute is embedded in mimesis, a prop is embedded in a game of make-believe.

Western cultures are sensitive to ‘idolatry,’ which in one sense is to substitute an idol for the divine. Given Gombrich’s theory, a cultural analogy can be made between an idol, which is a ‘fetish’ object, and a prop ([1963] 1978, p. 7). Something is said to be ‘fetishized’ when it becomes the object of displaced fantasies and desires, and calling something a ‘fetish’ in contemporary social criticism amounts to claiming that it has gained power over its maker (Grieve, 2001, pp. 120-121). The prop fits this mould by fulfilling displaced needs as an artifact bestowed with powers. Gombrich takes advantage of the cultural link between fetishization and substitution to argue that objects and bits of language are capable of serving this role when they “conjure up” what is needed of them in the right context ([1963] 1978, p. 5). As if in a séance, participants (e.g., readers) collaborate with conjurers (e.g., authors), intentionally or not (p. 10). After all, bits of language, too, can be fetishized or accused of idolatry. Still, substitutes can be more than mere replacements. They can fulfill a vacant need with something creative and unexpected.

It is not easy to pick up on Walton’s reliance on Gombrich without prior exposure to both theories. In analytic aesthetics, the dominant theoretical poles have been resemblance theories and symbolic conventionalism (Meynell, 2008, pp. 14-15). However, Walton’s adoption of the substitution model has enabled him to work with similarity relations, conventional traditions, and other factors in tandem. The theory of make-belief can deal with how a taxidermy bear resembles a bear, how a halo surrounding the head of a figure symbolizes a spiritual character, and how Anish Kapoor’s *Cloud Gate* does not resemble or symbolize anything in particular, simply by conceiving that the taxidermy bear, the halo and *Cloud Gate* are all props in specific games of make-believe.

5.3: Reader-Response as Make-Believe

Reader-response theory is a well-established school of literary criticism, yet it serves as the best template for understanding the theory of make-believe. As its name suggests, it is concerned with the way readers respond to literature and challenges the self-sufficiency of the static text (Fish, 1980, p. 2). Instead of an inquiry into meaning, it looks into what literature actually *does* (Iser, [1976] 1986, p. 360). It studies the ways in which a text acts upon its readers and the way its readers negotiate their reactions to the text, in a procedure that actualizes the work of literature (Fish, 1980, p. 3). For readers, this procedure involves "interpretive strategies" stemming from (constitutive) rules shared within their "interpretive communities" (p. 14). In some ways, the text is seen as an incomplete canvas that provides instructions for completion, performatively actualized by the reader according to shared conventions (Iser, [1976] 1986, pp. 367-370). It must be noted that, in this theory, the reactions of the reader are not considered to be mere responses to reasons, but the causal outcome of reacting to the text (Blackburn, 2005, p. 309). As such, the theory has epistemologically externalist leanings for being invested in the causality of the text and its readers' interpretive communities, with a reputation for flirting with all-out relativism.[27]

The philosophical program spawned by Walton fails to link his thought to reader-response theory.[28] However, although it regards reasons as response-provoking, the theory of make-believe does a more or less similar job and Walton borrows some of his

[27] "I am not claiming that there are no facts, I am merely raising a question as to their status: do they exist outside conventions of discourse (which are then more or less faithful to them) or do they follow from the assumptions embodied in those same conventions?" (Fish, 1976, p. 1017).

[28] The similarities between reader-response and Walton's theory run deep. In *Mimesis as Make-Believe* (1990), Walton spills enough ink on Stanley Fish, who is perhaps the most prominent figure associated with reader-response theory, to reveal that he is not oblivious to the details of this school of literary criticism (pp. 99-102).

technical vocabulary from this literature. There are at least two good reasons for drawing this connection. First, it helps to explain the epistemic strength of Walton's theory. For him, to engage with fiction is to engage in a game that involves (constitutive) rules[29] which categorically *prescribe, mandate or prompt* imaginings (Walton, 1990, pp. 38-40). Though left without a warning, these three verbs are not always interchangeable. Walton regularly uses 'prompt' to describe a causally induced (automatic and unreflective) response, and "prompter" to describe a prop that causes imaginings (pp. 21-43). In contrast, 'mandate' and 'prescribe' are often used in reference to fictional content that is supposed to be imagined. So, instead of focusing on how imagining is induced within a game of make-believe, Walton simply endorses appropriate prompts as reasonable mandates. For example, actually seeing the Statue of Liberty justifies me in fictionally seeing (i.e., imagining) a person bearing a torch because there is a rule to imagine what I am prompted to perceive from a statue. But, if in my ignorance, I found the torch to look like soft-serve ice cream, such similarity does not justify me to imagine that it is ice cream because of other reasons—cultural, factual or authorial—that prescribe me to imagine a torch, even if I am not aware of these reasons. In countervailing prompts by prescriptions, Walton balances the epistemic value of causes against reasons in his theory and rectifies the reader-response tendency to forfeit claims about the determinateness of interpretive content (i.e., objectivity). Even as socially, culturally or psychologically constructed (by the strategies of interpretive communities), the epistemic relativity of fiction does not follow its ontological relativity.

[29] Prompters, unlike prescriptions or mandates, cannot be anything but constitutive or categorical rules in games of make-believe.

The second reason for drawing a link to reader-response theory is to point out their core ontological correspondence over the conception of ‘representational objects’ which, I find, helps to elucidate the relationship between a participant and a prop in a game of make-believe. In broad strokes, reader-response theory argues that a literary work cannot be understood without considering the author’s plan (i.e., intent) in conjunction with the response of its situated readers (i.e., affect) which complete the act of reading (Felluga, 2015, p. 261). In finer detail, the theory views a *literary work* as the outcome of a three-step process: first, there is the *text* which is written by an author according to a plan; second, there is the *response* of the reader to the text according to their disposition; and, third, there is the range of possible *readings* culminating in the executed plan of the author (i.e., the voice of the narrator) received by the disposition of its readers. The literary work is completed at the stage of ‘reading’ (i.e., the interpretation). Authors tend to plan their text for a specific range of interpretative possibilities so that their project goes as intended. But, of course, they make mistakes or fail to plan properly. This is why the reader’s response must also be authoritative. The reception of a text is just as important as the aspiration of its author.

I see Walton’s theory as a more expansive version of reader-response theory. The complete ‘literary work’[30] is analogous to the ‘game of make-believe’; a particular ‘reading’ to a particular ‘imagining’; the ‘reader’ to the ‘participant’; and the ‘text’ to the ‘prop.’ The difference is in Walton’s wholesale enfranchisement of the representational arts. Just as in reader-response theory, a game of make-believe is the outcome of the author’s executed plan in conjunction with the response of situated participants, which

[30] Walton uses ‘work’ or ‘representational object’ more or less synonymously.

explains why Walton's understanding of fiction (a.k.a. 'Walt-fiction') only extends over literary works that prompt *imaginings* in *games of make-believe*—because, for him, all representational arts involve an analogous three-step process like the one considered.

Beyond this structural symmetry, I call attention to the process that *completes* or *actualizes* a literary work in reader-response theory. This is an ontological account, because the *complete* or *actualized* 'work' is not equivalent to the object called 'text,' though the latter is a necessary component of the former. The complete work is a composition and, so, failing in composure the work is in a state of incompleteness. Walton extends this ontological account over other representational objects, such as paintings, statues and toys, because they are also components in the actualization or completion of something more, say an exhibition of a painting or statue, or a game played with a hobby-horse. The representational object is a 'prop' that provides a work world, the compositing activity is a 'game' that generates a game world, and 'actualization' or 'completion' amounts to the 'make-believe' of a fictional world.

5.4: What is Make-Believe?

The hyphenated compound 'make-believe' does not appear in the overwhelming majority of reference sources in philosophy. So, what is it doing at the core of a philosophical theory? Make-believe is widely used to describe a way of imagining or pretending that something is the case.[31] But this is not exactly what Walton has in mind. For him, make-believe encompasses what usually counts as imagination or fantasy,

[31] Make-Believe (noun): "The action of making believe; pretence, fanciful imagining (esp. that things are better than they really are)" (Oxford English Dictionary, 2024); "a pretending that what is not real is real" (Merriam-Webster, 2024).

dreams and certain instances of perceptual experience, in addition to a certain way of entertaining thought.

A useful clue comes from the analogy between the *reading* of a *literary work* and the *imaginings* in a *game of make-believe*. Here, 'reading' refers to a justified interpretation of a text as the antonym of 'misreading.' It describes what could justifiably be *believed* about a literary work of fiction. Similarly, when it comes to imaginings, first, what could justifiably be imagined (i.e., justified true imaginings) in a game of make-believe is called 'authorized' imaginings and, second, there are perceptual and suppositional imaginings in make-believe.

To follow the first point, what distinguishes justifiable imaginings in a game of make-believe? When a participant engages in the make-believe of a thought experiment, *imaginings that generate the game world appropriately and as planned* are **authorized imaginings** and result in the generation of game worlds that are **authorized games** of make-believe (Walton, 1990, p. 51). Participants may imagine all sorts of things that were not meant to be imagined. If they engage in unauthorized imaginings, then they would be playing an unauthorized game. For example, I may imagine that medieval representations of angels in paintings depict extraterrestrial aliens of contemporary lore. But this would be a stretch, even though some works of fiction do authorize a wide range of imaginings (Meynell, 2014, p. 4160). In the case of the Bible, it may be authorized to imagine different characters with various hairstyles. However, to imagine them with baseball caps would be anachronistic and unauthorized even without any clear indication to the contrary. In every case, the difference between the work world and the game world is the difference between what there is in the fiction itself and what participants imagine there to be. So, a larger range of imaginings than would be authorized will, for participants, continue to be unauthorized.

Thought experiments often generate unanticipated responses. For instance, a philosopher could pose a thought experiment such as "Try to imagine something that is both a circle and a square at the same time," and conclude that "such an imagining is impossible." However, a participant could justifiably interpret the word "thing" as a three-dimensional object and respond by saying "I can easily imagine such a thing, and it is called a cylinder." In this case, even if the philosopher was only speaking about a two-dimensional "thing," the work world of the thought experiment is phrased such that it allows participants to imagine three-dimensional objects in their game worlds. The mismatch between the goals of the work world and the resulting game world is epistemically surprising for the philosopher because it shows how the thought experiment allows for more fictional truths than anticipated.

The fictional world of a thought experiment *becomes epistemically useful when the author manages to keep the limits of the authorized game within control, such that a match between the game world and the real world is achieved*. This match enables the participant to find imagined connections between beliefs and conceptions or to form useful conjectures or predictions about the real world. Consider the following thought experiment, called the *Paperclip Maximizer*:

> [Imagine] a well-meaning team of programmers make a big mistake in designing its goal system. This could result […] in a superintelligence whose top goal is the manufacturing of paperclips, with the consequence that it starts transforming first all of Earth and then increasing portions of

> space into paperclip manufacturing facilities. More subtly, it could result in a superintelligence realizing a state of affairs that we might now judge as desirable but which in fact turns out to be a false Utopia, in which things essential to human flourishing have been irreversibly lost. We need to be careful about what we wish for from a superintelligence, because we might get it. (Bostrom, Ethical Issues in Advanced Artificial Intelligence, 2009, pp. 380-381)

Here, Bostrom brings together two easily understood ideas to make a rudimentary prediction and a warning. Suppose that I think those with the best intentions and the best tools produce the best solutions which, then, brings me to look forward to the luxury of a fully automated artificial intelligence-guided future for humanity. But when I imagine the fictional world of the *Paperclip Maximizer*, I notice how these beliefs could plausibly combine to produce a dystopian future. Subsequently, I no longer blindly await a fully automated artificial intelligent guided future for humanity. I dread it. In such a case, the thought experiment enables imaginative engagement with a hypothesis. The game world allows the participant to assess if certain beliefs are in harmony or what could be anticipated from an event. Sometimes I have enough, or more than enough, information about the world than I realize. Thought experiments help me to make useful connections and notice interesting relationships with what I already believe. Furthermore, they help me to examine and challenge myself in fictional situations. Of course, make-believe may seem like a flimsy epistemic technique, but expectations must always be tempered in proportion to the reliability of a method.

To follow the second point, consider how ‘imagining’ usually refers to fantasy, but also connotes images. I can imagine having blue hair, in which case I would be generating mental imagery to consider, suppose or entertain how it would be to have blue hair. In such a case, I imagine something *perceptual*. On the other hand, I can very well imagine that my reflection in the mirror is not how I actually look like. This does not alter the image but only dissociates my identity from it. I simply entertain a thought, a hypothesis, or something *suppositional*, even if about an image.

Reading almost any work of fiction involves both perceptual and suppositional imaginings. This much seems obvious. But Walton’s efforts in making room for the representational arts by over-stretching what it means to ‘imagine’ conflates the difference more than usual. In his account, I do not only imagine what it would be for me to have blue hair when I imagine myself with blue hair. I also imagine myself as such in response to actually looking at a picture of myself with blue hair. This is most unusual because ordinarily, I would consider such a description to be the outcome of a linguistic mistake. Yet, for Walton, gazing at a picture or imagining a picture have the act of imagining in common, even if in the former the involvement goes unacknowledged. This way of understanding imaginings marks an unprecedented phenomenological blend of imagination and perception that risks conceptual ambiguity about the two (Guter, 2010, pp. 121-122). Still, I think the risk can be mitigated. Consider how looking at a painting of a mountain prompts certain imaginings in response to a painted canvas just as reading a text about a mountain prescribes certain imaginings in response to bits of language. The painting is a perceptual aid (i.e., a prop/prompter) to make-believe the seeing of a mountain, but it would be a mistake to think of it as a way of actually seeing a mountain since the painting is not a mountain. One sees the painted surface in the right condition and is prompted to imagine a mountain. In this respect, the disambiguating technique is to point out that the perceiving of representational objects is ontologically prior to the imaginings prompted by them.[32]

This technique is limited. When it comes to thought experiments, the ambiguity is not due to a phenomenological blend of imagination and perception, but a blend of suppositional imagination (i.e., imagining that/the pretense to believe) and perceptual imagination (i.e., mental imagery/envisioning). Most games of make-believe involve both, but sometimes they come apart in important ways without notice. This ambiguity is perhaps the foremost instigator of confusion in thought experiments, especially because imaginability (as conceivability) is often assumed to be a good test of possibility. Walton touches on this issue and takes it to be an obvious fact that fiction can prescribe impossibilities and participants can imagine certain contradictions, such as paradoxes and false perspectives (1990, pp. 63-67). But, since thought experiments are epistemically more demanding than other forms of fiction, some clarification is needed.

First, it should be remembered that thought experiments involve a blend of suppositional and perceptual imaginings. If a mathematician taught me that the so-called ‘imaginary numbers’ are lateral units[33] to real numbers, extended on a perpendicular axis to the real number system, I can *imagine that* their magnitude extends in a new dimension. This allows me to *envision* them graphed, which I could not have otherwise

[32] Stated differently, the representational object (i.e., the prop) is the object of intentionality, but the object represented or presented by the content (e.g., the mountain) is the intentional object. Notwithstanding reflexive prompters (e.g., dolls and statues), intentional objects are not identical to the objects of intentionality.

[33] Johann Carl Friedrich Gauss, the renowned mathematician, is quoted as having said: “If for instance, +1, -1, √-1 had been called direct, inverse, and lateral units, instead of positive, negative, and imaginary (or even impossible) such an obscurity would have been out of question” (Moritz, 1914, p. 282).

conceived of doing. Here, suppositional imagining leads to perceptual imagining, I have learned something in the process and there is no risk of ambiguity. In contrast, take Walton's example of "an elf who squares the circle" (1990, p. 64). I can *envision* a humanoid that I *suppose* is capable of doing something impossible, but, since the technique of squaring the circle cannot be imagined in any detail, I cannot *envision* how he does what he *supposedly* is capable of doing. Here, suppositional imagining does not lead to perceptual imagining,[34] I have not learned anything and there is a risk of ambiguity if the difference between the two ways of imagining is ignored.

It is crucial, in the second place, to not confuse the content of imagination as something more than make-believe. Neither suppositions nor mental imagery of fictional possibilities are guaranteed their complimentary match in reality. It seems obvious that just because I can imagine Peter Pan losing his shadow (suppositionally and perceptually), it does not follow that anyone or anything in the real world could possibly lose a shadow. But this is not always so blatantly obvious. For example, both Gospels of Mark and Mathew suggest that salt could lose its savour (The King James Bible, [1769]1989, Mark 9:50; Mathew 5:13). Although neither case presents a thought experiment about salt, it is fair to wonder if salt, as Sodium Chloride, could possibly lose its savour under normal conditions without impairment to one's sensory capacity. I can easily make-believe a scenario where I pour some table salt on my tongue only to sense its granular texture without the usual saltiness. If I do, my imagining would be both suppositional and perceptual, but would it also suggest that it is really possible to have such an experience in the real world? My answer is negative because perceiving salt as

[34] 'Elf(x)' is perceptual, but 'SquaresTheCircle(x)' is suppositional, if the statement is regimented as:

$$\exists x\ (\mathrm{Elf}(x) \land \mathrm{SquaresTheCircle}(x))$$

savoury is identical to the sensory detection of Sodium Chloride under normal conditions. Unfortunately, I suspect that many of those committed to the use of thought experiments in metaphysics and beyond would not find this point so obvious. Thus, I believe the mistake of moving from what could fictionally be true to what could possibly be true—a move from the fictional world to the real world—is the root of many mistakes and confusions in using thought experiments.

5.5: Walton's Machinery of Generation

Make-believe cranks out fictional truths by utilizing various gears and levers, metaphorically speaking. Walton calls these *components of the system generating fictional truths* the '**Machinery of Generation**' (1990, pp. 139-140). Here, I introduce the Machinery of Generation at work in thought experiments.

A game of make-believe is generated by the imaginings of participants. Walton's theory holds that the work of fiction relies on '**Principles of Generation**' to *determine fictional content and guide participants in their imaginings* (1990, p. 38). Principles of Generation determine the fictional content of a work's game of make-believe by constituting conditional prescriptions, or sets of rules, for justifiable imaginings (p. 41). They do not ensure the intentions of the author nor the conditions of the fictional world's possibility. Instead, Principles of Generation guide participant engagement.

The work of fiction offers one class of principles by itself. For example, a narrative guides the reader through sentences that are read and imagined. If I read in a narrative "This morning, mother died," I imagine what it explicitly states, that someone's mother died this morning. This type of provision is directly made by the author and it accounts for the '**direct**' Principles of Generation (Walton, 1990, pp. 140-144). They are *provided by the work world and describe the ways authors stipulate Principles of Generation*.

Most of the time, fiction relies on *indirect ways to guide imaginings*. **Indirect** Principles of Generation are *provided by participants* (Walton, 1990, pp. 140-144). Most Principles of Generation function indirectly, such as general conventions, background beliefs, tacit rules, common psychological and perceptual habits, capacities and constraints of participants (Meynell, 2014, p. 4159). For instance, "Once upon a time, in a land far faraway…" indicates that a fairy tale of some kind is about to begin because fairy tales usually begin with such a motif. In this case, this is a conventional cue, but in other cases, it could be based on a variety of other mechanisms. Whatever the case may be, as the contributions of participants, indirect Principles of Generation do not depend on the work alone.

In total, the work of fiction together with the Principles of Generation ground the content of make-believe. The failure of participants to understand or interpret the work properly, by imagining inappropriate or unfitting content, means they lack the needed Principles of Generation or are employing them in the wrong way so that they do not properly condition imagination (Meynell, 2014, p. 4159). Principles of Generation do a lot of heavy lifting and, because of this, they are central to understanding imaginative engagement with any work of fiction.

Walton specifies two important and all-purpose principles of indirect generation by which common tendencies of implication can be characterized and governed when it comes to questionable or indeterminate fictional truths. These are the '*Reality Principle*' and the '*Mutual Belief Principle*.'

The **Reality Principle** guides the participant to *interpret the fictional world like the real world as much as possible* (Walton, 1990, p. 147). It explains that there are implicit fictional truths in any make-believe that match how the real world actually happens to be. This allows participants to supplement much of the rest, as needed, by simply appealing to what is known about the real world. For example, assuming that, in some fictional scenario, a human has red blood, even with the absence of any clue from the work world, is justifiable according to the Reality Principle. In a fictional world, a human has red blood simply because in reality humans do have red blood. This would even be true in a fairy-tale if there are no other Principles of Generation to the contrary.

On the other hand, the **Mutual Belief Principle** guides the participant to *interpret the fictional world through the beliefs of the work's author, the community of belief the author belongs to, or the beliefs of the work's intended audience*. Even when there is overlap, the beliefs of others are to be considered independently of the participant's own beliefs and epistemic commitments (Walton, 1990, p. 152). The idea is that there are implicit fictional truths which depend on the beliefs of the author or the relevant community of belief in any imagined scenario. For example, to be able to sufficiently appreciate the epic of Gilgamesh, enough familiarity with the life and mythology of the ancient world is required. Different deities from the Mesopotamian pantheon play a part, and interpreting the story requires some familiarity with the relevant cultural background. Therefore, imaginative engagement with fiction may call for beliefs and assumptions unfamiliar to the participant. Because of this, the participant needs to consider what interpretive tool is needed, use the appropriate one for the job, and learn enough about it. However, this is easier said than done. Interpretations can and do vary, often enough for good reason, the right information is not always available, and the most attentive participants make mistakes. The best thing to do, here, is to rely on the Mutual Belief Principle because it brings to bear a kind of tolerance and pluralism regarding background assumptions (Meynell, 2014, p. 4161). Imaginings should be expected to be heterogeneous in practice.

In "Race, Ethnicity, Biology, Culture" (1999) Philip Kitcher retells a thought experiment from E. O. Wilson and J. Philippe Rushton, which I shall refer to as the *Martian Naturalist*:

> Imagine a Martian naturalist visiting [Earth] for the first time and observing our species. What infraspecific divisions, if any, would the Martian draw? Rushton announces confidently that they would spot three geographical 'races' with different body types. But simply noticing the phenotypic variation in height, bone thickness, skin colour, or whatever should not inspire the Martians to divide out species into races—Rushton's Martians (and probably Rushton himself) make a mistake against which Ernst Mayr has inveighed for so long that it has become part of the standard equipment of any field naturalist concerned to identify the species in a particular area. Only the uninformed rush in and divide sexually reproducing organisms according to the differences that strike them, the outsiders, as salient. […] Taxonomic divisions should be grounded in distinctions that the organisms themselves make, in the propensities for mating and reproduction. […] So, a Mayrian Martian, looming at our species, would attend, above all, to the facets of our

> reproductive behaviour, noting not simply the phenotypic differences but seeing that in some locales, like the United States, those phenotypic differences correlate quite strikingly with mating patterns. To return to our fantasy and state the moral more soberly, intermarriage statistics are crucial because those statistics (poor though they are) are proxies for what is biologically crucial in making taxonomic divisions. (Kitcher, Race, Ethnicity, Biology, Culture, 2003, p. 241)

The thought experiment imagines what racial divisions a Martian naturalist visiting Earth for the first time would draw if the Martian were to draw any divisions. But what the Martian would find striking about humans depends on how participants imagine the Martian to be in the first place: are they physiologically humanoid (e.g., the little green men or the greys) or are they similar to cephalopods (e.g., the Heptapods from the (2016) film, *Arrival*)? The former would likely not find the fact that humans are bipeds most striking, while the latter possibly would. Similar worries result from a Martian racial classification of humans. Participants could fail to identify how fictional truths can be interpreted differently based on the background assumptions at play, where the Mutual Belief Principle takes account of the relevant background beliefs that are not explicitly provided by the thought experiment.

Scientists and philosophers make up their own particular cultural communities and bring implications of their particular subdisciplines to bear in make-believe. Different training, conventions, mental tendencies and dispositions inform one's imaginings (Meynell, 2014, p. 4161). So, shared mutual beliefs make a significant difference in imaginative tendencies. Because different communities rely on their own mutual beliefs to interpret fiction, it is fair to expect that the Mutual Belief Principle would allow the same fiction to be interpreted differently by dissimilar communities. Again, a certain diversity of opinion needs to be tolerated in make-believe. This incessant possibility for pluralism means that in some cases the content of a fiction's imaginings, dependent on the Mutual Belief Principle, is ambiguous (p. 4161). There could be no determinate outcome for some thought experiments. In such cases, Principles of Generation equally give rise to incompatible games of make-believe, and ample attention needs to be given to practices of interpretation.

One last thing to keep in mind is that drawing a clear line between the Reality Principle and the Mutual Belief Principle may not always be an easy or achievable task.[35] The Reality Principle points to what the participant believes about the real world, while the Mutual Belief Principle is about what the participant does, or should, know about the beliefs of others. This suggests that the Mutual Belief Principle is a subset of the Reality Principle. However, for Walton, there can be uncertainty and disagreement about which principles are applicable in some cases (1990, p. 138). Principles of Generation are the reconstruction of judgments about what is fictional, rather than stand-alone rules for making judgements (p. 185). These principles can point in contrary ways to answer to different needs in different cases, and formulating universal or systematic meta-principles for determining all Principles of Generation does not seem promising for Walton (p. 169). This is because these principles depend on the imaginative tendencies of

[35] Why should such a seemingly straightforward distinction be susceptible to such confusion? I can suggest a possible connection. Walton more or less borrows the two principles above from David Lewis's (1978) paper, "Truth in Fiction." Lewis, however, approaches fiction in a very different way, by relying on the metaphysics of possible worlds. The Principles of Generation are adopted in close approximation to distinctions made by Lewis (pp. 44-45). Walton imposes prescriptive force upon these distinctions and stretches the use of 'belief' to include conventions, impulses, habits and so on, but his efforts only go so far, becoming the cause of unnecessary complexity.

participants, and the ability to neatly pick apart these tendencies will be limited. Still, I do not see why Walton should dissuade anyone from trying to use Principles of Generation for making appropriate judgements in participating, authoring or critiquing thought experiments.

As mentioned, implication is (usually) accomplished by either the Reality Principle's reliance on beliefs about the real world or the Mutual Belief Principle reliance on beliefs about the social world. Now, if the social world is embedded in the real world, then the Mutual Belief Principle must be a subset of the Reality Principle. Instead of being about the participant's beliefs, the Mutual Belief Principle is about the beliefs of the fiction's author or the work's relevant community of belief. Yet, beliefs about what others believe are sustained by beliefs about the world. Also, a significant portion of one's beliefs about the world are accepted on trust from others (i.e., the opinion of experts, family, friends, etc.). One's beliefs about the world cannot easily be sorted into those about the real world and those about the social world. One only learns about the real world through the social world (e.g., experts), and vice versa (e.g., experience). Given any belief about the real world, if it is not one that I hold, but one that I acknowledge others do, I hold it as a belief about the social world. For example, my belief about the endangered status of sea turtles is a belief about the real world. But if I were to doubt it for any reason, it remains my belief that the experts and a segment of the public still believe that sea turtles are endangered. This way, the 'real world' would simply be the world as I know it to be. For the theory of make-believe, the upshot is that the Mutual Belief Principle cannot be the only principle that ensures a level of tolerance and pluralism. The Reality Principle has that capacity, too, by recognizing that beliefs about the real world may vary from one individual to another, not just one community to another. Thus, the way that fictional worlds match the real world turns out to be a private affair, because the Reality Principle is, at least in some way, subjective given that it guides each participant to imagine according to what they believe about the real world, not according to the real world itself.

The clarification above leads to another related issue. Participants often have to settle on a strategy without being sure of the principles at play. Their strategy must be to compare the possible implications of the Reality Principle (i.e., what they believe) with those of the Mutual Belief Principle (i.e., what they could come to believe). Walton covers a case where two individuals, Loretta and Mabel, are reading a story about an anti-social character named ‘Andy’ (1990, p. 159). Loretta relies on the Reality Principle and imagines that fictionally Andy suffers from a neurological disorder since his behaviour would be indicative of such a disorder in a real person. Mabel, on the other hand, relies on the Mutual Belief Principle and imagines that Andy is possessed by the devil since such behaviour is indicative of possession by the devil in the author’s society. Both Loretta and Mabel verify their beliefs by researching the relevant scientific and historical resources. Loretta looks into the science of psychology and Mabel looks into the history of beliefs about possession in the author’s society. But each case reaffirms the initial suspicion. In such cases, Walton argues the Mutual Belief Principle should be “understood as the determiner of what fictional truths are implied” to best facilitate the author’s creative project. So, Loretta is unnecessarily transposing her beliefs about the real world into the story, and she should rethink her reliance on the Reality Principle.[36]

[36] This example demonstrates how Walton's theory could easily accommodate considerations regarding reflective equilibrium in thought experimentation without sacrificing its own methodological commitments.

Therefore, the normative priority should be placed on the Mutual Belief Principle in most cases.

In thought experiments, should the normative priority be placed on the Mutual Belief Principle instead of the Reality Principle? I expect that to be a mistake. Most thought experiments mandate participants to rely on the Reality Principle because the individual response of the participant is sought after. Sometimes thought experiments involve *de se* imaginings[37] such that participants need to imagine as someone that they are not, in the first-person perspective, so that they can imagine situations, events or experiences they could not or would not otherwise imagine (Walton, 1990, p. 34). This way of imagining is achieved through the act of *pretense*[38] by role-playing what someone else is to imagine (p. 35). Thus, such cases prioritize the Mutual Belief Principle. However, it may be safe to say that these are not the majority of thought experiments. The epistemic value of thought experiments mostly rests on the participants responding as themselves. Thought experiments are not ordinary fiction. They serve epistemic needs for which they deserve to be paid particular attention. Indirect Principles of Generation, especially the two principles above, do sensitive work in this context. It must be remembered that the Mutual Belief Principle does not override the Reality Principle in the way it can for literary fiction. The Mutual Belief Principle only takes priority when

[37] Walton calls “imagining *de se*” the act of imagining oneself having experiences as someone else and in the first-person perspective, or the act of imagining being someone else or something else from the inside (1990, pp. 28-35). “These self-imaginings are important” for Walton “even when our main objective is to gain insight into others. In order to understand how minorities feel about being discriminated against, one should imagine not just instances of discrimination but instances of discrimination against oneself; one should imagine experiencing discrimination. It is when I imagine myself in another’s shoes (whether or not I imagine being him) that my imagination helps me to understand him” (p. 34).

[38] Broadly speaking, *pretense* describes the performance of all of the acts in a game of make-believe required from a participant (Walton, 1990, pp. 379, 391).

participants need to know about the experience of others, as they would when *de se* imaginings are mandated.

Above, I mentioned how in Walton's example both Loretta and Mabel verified their beliefs through research. But, should participants be expected to conduct extensive research to properly engage with a thought experiment? I think that would be a tall order. Not everyone can engage with every thought experiment. Thought experiments are embedded in particular contexts of use, in particular arguments, and require participants to keenly follow their mandates in generating imaginings. For instance, thought experiments focused on *de se* imaginings will be more demanding. The participant needs to be familiar with the community of belief in question and try harder to imagine being someone else. However, how familiar participants need to be in each case is not an easy question to answer. There needs to be a reasonable threshold that does not demand extra-textual research for the average participant. I cannot say more about such a threshold, nonetheless, it is crucial to pay attention to it in any thought experiment.

There is an affinity between thought experiments and forms of fantastic fiction, particularly the genre of science fiction. Science-fiction usually offers speculative trajectories of how some society prescribes imaginings about its potential futures (Luckhurst, 2011, p. 1257). Thought experiments have similar enough of a job, but they have to carefully balance their Principles of Generation to track something epistemically useful. In both forms of fiction, the Reality Principle cannot correct for unrealistic fictional truths postulated in the narrative. It just fills in for whatever is not stipulated directly. The rest of the make-believe is not guaranteed to live up to realistic expectations. If imaginings are generated according to our conceptions of the real world, they could become misleading if they then depart from the contingent boundaries that make them empirically relevant (Hardimon, 2017, pp. 45-47). If the concept shapes beliefs about the real world without also being conditioned by it, it will not have the epistemic import that it would have had if it were conditioned by empirical reality. There is a threshold for empirical relevance. Technical or contextual uses of concepts are not normally corrected by their non-technical, non-contextual use. It tends to be the other way around: non-technical or non-contextual uses of concepts are informed by their technical or contextual use. Otherwise, thought experiments would be mere fantasy, unimpeded by scientifically unexplainable content (Rankin, 2011, p. 1054). Fantasy may be epistemically valuable, but it does not have to be.

To learn about the world itself, one needs to primarily rely on the Reality Principle. The standard for the Reality Principle is *what one believes about reality*, not reality itself. When there is a deficiency of facts known about the real world, one can try to learn more before going any further. Sometimes, this means looking into the facts, conducting research, taking a good look and checking to see what the real world is like. Other times, one must learn about unfamiliar people, cultures, worldviews, competing positions and expert opinions. The Reality Principle captures everything I believe about the real world, including what I believe about the real communities of belief in the Mutual Belief Principle. In contrast, the Mutual Belief Principle is about *what others believe about the real world*. A thought experiment has considerable wiggle room when generated under this principle since its work world is incomplete and needs participation to determine results; so prior assumptions and beliefs are going to be consequential (Meynell, 2014, p. 4163). One simply needs to articulate the relevant Principles of

Generation at play and assess their intended force and consequence for the purposes at hand (p. 4167). As simple as it may be, this is a framework for detecting the cause of major disagreements over thought experiments.

5.5.1: Revised Machinery of Generation

Earlier, I introduced how Principles of Generation come in direct and indirect varieties. Of these, the indirect principles account for the majority of the Principles of Generation. They are, in effect, the contribution of participants, instead of the work of fiction itself, and could best be understood as the reconstruction of the general ways in which participants engage with fiction (Walton, 1990, p. 185). Make-believe relies on a complicated, shifting and often competing set of tendencies of understanding, precedence, convention and salience (p. 169). To clearly understand what features of a work generate which fictional truths, the relevant features must be controlled by the author and fully ascertained by participants in a mutual project of understanding (p. 165). The trouble is theoretically capturing this shared understanding. It is not going to be easy and, as Walton argues, indirect generation can be disorderly (p. 169). This being said, I take it that sometimes Walton is not trying hard enough to clarify his theory by finding comfort in the admission that things are complicated. Things are complicated, but that is no reason to settle for less.

Are all indirect principles subsumed under the principles mentioned? Walton's language suggests as much (1990, p. 140). However, he does note that a "diversity of devices" is commonly found to do the job of implication (p. 169). Moreover, he acknowledges that indirect Principles of Generation include devices that are not "reasonable grounds" for justifying an implication (pp. 167-168). He makes this point evident in the following:

> If even the flimsiest evidence relation can ground implications, provided it is reasonably conspicuous, one should expect there to be implications involving no evidence relation at all (neither actual nor believed), but merely a sufficiently salient connection or association of some other sort. There are such, of course. Sometimes the association holds only within a given representational tradition, and sometimes it is established by representations themselves. (Walton, Mimesis as Make-Believe, 1990, p. 165)

In other words, the principles articulated are but a subset of indirect Principles of Generation. The Reality Principle and the Mutual Belief Principle are **Principles of Implication** because they are the *subset of indirect Principles of Generation grounded by evidence*. Other types of indirect generation are left uncategorized by Walton. So, in connection to my earlier use of tacit dispositional strategies in practical circumstances, I call such non-implicit principles '**tacit**' Principles of Generation for being *the subset of indirect Principles of Generation that are grounded by mere salience*.

5.6: Walton's Mechanics of Generation

At this point, I have introduced the main features of the theory in particular reference to thought experiments, but a general distinction needs to be made. What has thus far been introduced is only the Machinery of Generation. But the gears and levers of the Machinery need to be organized and set in motion to generate a dynamic game of make-believe. To account for the *operations of the Machinery of Generation*, Walton coins the '**Mechanics of Generation**' (1990, p. 140). Once the gears and levers of the

Machinery are identified in a particular game of make-believe, the inspection of their operation becomes the job of the Mechanics of Generation.

The prop and the principles are the Machinery generating fictional content. The majority of fictional content, especially when generated according to Principles of Implication, is the cognitive output of participants in following prompts or direct fictional truths to determine what indirect fictional truths are prescribed to be imagined. However, it is important to recall that ‘fictional propositions’ are only heuristic devices. Since implication is, and inference is not, a relation between statements, and ‘fictional truths’ are not like truths of true statements, but *prescriptions*, *mandates* or *prompts* to imagine fictional content, Walton’s ‘Principles of Implication’ is a misnomer. No relation is implied between literal antecedents and literal consequents when indirect fictional truths follow imaginings generated by prompts or direct fictional truths, unless reconstructed for the sake of clarification. Thus, ‘Principles of Implication’ are best understood as ‘Principles of Inference’ because they guide participants to imagine indirect fictional truths as a consequence of prompts or direct Principles of Generation. Walton’s ‘Mechanics’ is a mere suggestion. But, since rules of reasoning are of concern in thought experiments, clarification is needed.

5.6.1: Revised Mechanics of Generation

Thought experiments can make arguments more forceful when they contribute persuasive power. To do so, thought experiments need to be dynamic enough to stimulate participation in support of the inferential moves of the argument. What I mean by ‘inferential moves’ is nothing more than the drawing of conclusions from a set of premises or assumptions which, in the context of make-believe, refers to the fictional truths of a thought experiment. Of course, determining fictional content rests on Principles of Generation, from direct stipulation to indirect implication and beyond. However, as Walton contends, the Machinery of Generation can barely hang together (1990, p. 183). As I maintain, in contrast to reader-response theory, the theory of make-believe accepts from participants both reasons and causes as appropriate responses. Consequently, there is no guarantee that investigation into the Mechanics of Generation would provide a satisfactory account of inference in every case, because compelling causal connections may not always translate into logical ones. As Walton insists, deciding what is fictional "must be sensitive to feedback from one's overall assessment of the work" (p. 184). After all, in the theory of make-believe, thought experiments are taken to be fiction, not argumentation. Thus, mechanically inspecting the Machinery only attempts to assess forces that are transmitted in support of an argument, without guarantees.

How can the Mechanics of Generation help the study of thought experiments? Inferences in make-believe often mirror inferences drawn in real life, even though the Principles of Implication take prominence in fiction (Walton, 1990, p. 359). In this context, mechanical inspection mostly means an audit of the chain of reasoning. In thought experiments, the concern is usually about the internal organization of components enabling the move from one fictional truth to another. I call this an '**endogenous inference**' for *moving from a fictional truth to another fictional truth within a thought experiment*. However, when a thought experiment is embedded in an argument, it is outwardly synchronized to the movements of that argument. I call this an ‘**exogenous inference**’ for *moving from a premise in an argument to a fictional truth within a thought experiment*.

Erroneous inferential moves could become problematic for thought experiments and, if they do, the responsibility could only be traced back to the author. Thought experiments are not used in open-ended ways in arguments, but in ways that support the author’s convictions. This is why authors conclude their thought experiments with what they think should result from participation. However, the fictional truths of a thought experiment are public and intersubjectively accessible (Meynell, 2014, p. 4156). So, if a purported result does not follow from careful participation, it can be shown to be mistaken. *A mistake by an inference endogenous to a thought experiment* is what I call an ‘**endogenous problem**,’ whereas *a mistake by an argument’s inference exogenous to a thought experiment* is an ‘**exogenous problem**.’

If endogenous problems arise due to a break of internal flow in thought experiments, exogenous problems are caused by their contextual incompatibility. Sometimes a thought experiment does not fit the argument employing it, even if it appears as if it does. This means that otherwise competent participants could find it difficult to imagine what it prescribes—also known as ‘imaginative resistance’[39] (Gendler & Liao, 2016, p. 405). It may be true that, in such cases, isolation from the context of use could ease the participant’s cognitive burden, but only at the cost of missing further

[39] Walton’s remarks on the problem of imaginative resistance are focused on cases involving moral deviation, whereby participants feel that they need to set aside their moral convictions “for the limited purposes of understanding and appreciating” a story (1990, p. 155). He suggests that participants are less willing to allow fictional worlds to morally deviate from the real world (Walton & Tanner, 1994, p. 35). This may be true in some ways, but Walton is missing the deeper issue. The reason why morally deviant fictions are known to cause imaginative resistance is not the misalignment of real world and fictional world morality. It is because participants continue to be committed to exogenous beliefs and values in make-believe, which is easily explained as exogenous problems in the context of fiction.

constraints set by the argument from the outset. Other times, an argument outright misleads participants. For example, participants can be distracted by ‘intuitive’ or dispositional responses and fail to see what else deserves attention. Whatever the case may be, a thought experiment cannot support an incompatible argument.

Chapter 6: The Method of Make-Believe

The main argument of this thesis targets the use of thought experiments in the metaphysics of race. My comparative analysis of the debate in Chapter Four has already set the stage for addressing the methodological disagreement in this chapter. Yet, the arguments offered by each philosopher could, at least to some extent, stand apart from the thought experiments in question and be defended separately, even if at a cost to their persuasiveness. So, before getting to case studies, it must be noted that unlike Glasgow, Jeffers and occasionally Hardimon, Haslanger and Spencer do not use thought experiments. Spencer is an ardent methodological critic of thought experimentation and Hardimon takes issue with the way certain thought experiments are conducted. This is not to say that Jeffers and Haslanger do not criticize the use of particular thought experiments, but that they do so short of a methodological criticism. For my part, I apply the theory of make-believe, detailed in the previous chapter, to case studies listed in Chapter Two.

6.1: Issues with the Machinery of Generation

6.1.1: The Reality Principle

6.1.1.1: The Dalai Lama. As mentioned in the exposition of the *Dalai Lama* thought experiment, Jeffers criticizes Glasgow for overlooking possible, and perhaps more intuitive, results. Jeffers accepts the possibility of someone saying that "we are all the same race," but he also imagines the possibility of saying that "we appear to be the same but this is only true at a superficial level—it remains the case that we are divided into races and our habits of association and our continued reproduction along these particular ancestral lines demonstrate [as much]" (2019b, p. 184). Yet, I believe that

Jeffers' second response could be shown to be a version of his first response. I can imagine in the *Dalai Lama* scenario that, just as Glasgow would have it, someone saying "We are all the same race," but I can also imagine, thanks to Jeffers, that the same person continues to add "…on the face of it when we don't know enough about each other." This is because, in this thought experiment, once individuals learn enough about each other's ancestry, particularly about what Glasgow calls the "geographical and cultural forces (such as language, dress, and popular places to find potential mates) [that] impact reproductive choices [and preserve] genealogical lines," they can come to know races in their new predicament, if not in connection to the history of the concept (2019b, p. 121). The *Dalai Lama* humans may look identical to one another, but they only need to communicate and observe social relations amongst themselves to have access to a wealth of information. The thought experiment leaves enough room for participants to imagine such information to be available in their game world. This time, the issue is not related to the Mutual Belief Principle. Glasgow has overlooked something about how his fictional world is supposed to work. He has underestimated the epistemic role of history and ancestry, together with the geographical and cultural forces that he acknowledges to be at play according to the Reality Principle.

The *Dalai Lama* thought experiment is an account of activists trying to solve the issue of racism in the context of police brutality and discrimination. In this fictional scenario, does the chemical agent in the water supply also redistribute the population in a homogenized or randomized pattern of residence? It does not. Racial populations are concentrated in patterns of racial segregation observable across the North American landscape. The police are well aware of these patterns, and it is not inaccurate to say that they take it to be their job to know these patterns. Does the chemical agent erase identification files and profiling data from police archives and servers? It does not. The police continue to identify and profile as before. Does the chemical agent eliminate family relations and social or community associations? It does not. But race often tracks these relations and associations. What about changing the political and economic motivations of the police for targeting certain populations? Again, the answer is negative. It would be absurd to think any of these are resolved by the chemical agent in question. Still, these are all important parts of the puzzle when it comes to how and why, in reality, the police actually operate as they do. The police target different neighbourhoods differently. The police use coercive identification techniques and gather and use personal information without public oversight. The social, political and economic environments inform the behaviour of the police in general, and techniques of identification inform their actions in particular. Between this and the persistence of family relations and social associations, I cannot conclude that the police would cease their racial policies entirely. If I have to appeal to what I know about the world to imagine this fictional scenario, then what I know about the real world allows me to articulate how, according to the Reality Principle, Glasgow cannot conclude the way that he does. He is forgetting about the implicit fictional truths that depend on the real world.

6.1.1.2: The Twin-Earth. The same issue is pointed out by Hardimon’s criticism of the *Twin-Earth* thought experiment. Recall that in this case, God creates a copy of the Earth which only now comes into being and, consequently, has no prior history. The point about history—a missing past—entails that inhabitants of *Twin-Earth* are without ancestry.

Glasgow wants to say that if participants can imagine them without ancestry, yet continue to divide them racially, ancestry is not essential to the concept of race. However, for Hardimon, if participants were to imagine their fictional doppelgangers, as Glasgow's direct Principle of Generation mandates, the doppelgangers would look like exact copies and their offspring would have to be imagined as inheriting some of these identical features, just as they would in the real world (2017, p. 45). Hardimon points out that Glasgow wants the doppelgangers to "look exactly like us" by imagining in accordance with the Reality Principle, at the same time as he anticipates the abandonment of the Reality Principle (p. 45). This is an inherent contradiction. If doppelgangers look identical, then at least some inheritable phenotypes are to be expected because, supposedly, the same biological and genetic processes are working on both planets. Nothing mandates otherwise. So, relying on the Reality Principle allows a participant to imagine that there would not only be similar lines of ancestry for the upcoming generations but also to imagine that the resemblance of "all the empirical evidence" guarantees the theoretical reconstruction of ancestral lines even in the absence of any ancestral precedence. A missing past does not prevent anyone from modelling and historicizing a past. Also, it is hard to imagine how such doppelgangers relate to their missing past, or whether they are meant to be aware of their missing past in the first place. This could potentially be a problem, but at least it is clear that participants are meant to imagine their doppelgangers to be making history exactly as they would themselves.

Glasgow could add that *Twin-Earth* does not have the same underlying biology. But that would misalign the fiction from reality such that it loses the epistemic import of his thought experiment. In such a case, drifting further away from the Reality Principle is not a solution.

6.1.1.3: George's Appearance-Transformation Machine. This is a good place to be reminded of how empirical concepts relate to empirical facts in the real world. Hardimon believes that in *George's Appearance-Transformation Machine*, the thought experiment misfires because it does not attend to how empirical concepts are connected with the contingencies of the world that they are about. Divorced from their relationship, concepts can lose empirical relevance.

The fondness for science-fiction themes is evident in Glasgow's narratives. He could stipulate magical potions and the like, but he chooses to go with things such as chemical agents. But this is not enough to guard against the empirical mismatch between the real world and a thought experiment's game world, simply because the Reality Principle can only do so much. It cannot fill in for imaginings that go beyond the purview of the participant's beliefs about the real world. It cannot provide anything near the sufficient range of physical or metaphysical modality needed for all imaginable fictional technologies. Science-fiction themes in thought experiments allow participants from different communities of belief to imagine shared scientific speculations. Once the Reality Principle is satisfied in some science-fiction, as much as it can be, the authorized game of make-believe allows for additional speculation according to indirect Principles of Generation. But whatever these additions may be, they remain tentative possibilities based on the speculations of the work world, rather than possibilities in the real world.

The tentative suggestion of a possibility is far from its empirical assertion. For this very reason, the empirical import of a thought experiment should never be over-estimated.

6.1.2: The Mutual Belief Principle

Quayshawn Spencer's criticisms of Glasgow's thought experiments are focused on the treatment of the Mutual Belief Principle. For Spencer, the dominant meaning of 'race' in any large enough linguistic community cannot be reliably intuited from the armchair, no less because there may not be a singular dominant meaning to isolate (2019b, pp. 234-237). It would be hard enough to map out the ordinary use of race talk and race thinking in American English, but it would be even harder to imagine any particular response to a thought experiment that would be representative of this diverse linguistic community at large. In such a project one supposedly needs to make-believe in an ordinary way. To do so, one needs to learn about the statistical trends because, given the size and diversity of the linguistic community in question, the ordinary uses will not be univocal.

Race is a polysemous word and a variable concept. So, Glasgow must include caveats concerning different ordinary uses of race and what they suggest about his argument and thought experiments. As mentioned before, some fictional worlds allow for more than one authorized game world. The Mutual Belief Principle calls for a kind of tolerance and pluralism over the background beliefs of different communities when it is the assumptions of others that are in question. For Glasgow, the background that needs to be taken into account is too diverse to determine a unified ordinary meaning. Consequently, there are different ways for participants to make-believe and, there is no guarantee that there will be compatible results to suggest a unified account of the ordinary concept of race.

Thought experimentation does not by itself reveal ordinary usage. There is a clear difference between how anyone employs a term or a concept in ordinary cases, and how a term or a concept is ordinarily used in a population. Thought experiments examine the former. The response of a participant only shows something about the way that they use the term or concept in make-believe, not how the concept or term is used by a population. It is true that experimental methods, such as those employed by Glasgow, can sample the population's responses to thought experiments and enable the estimation of the response of the entire population. However, since thought experiments only elicit imagined responses, and Glasgow's are extraordinary fictional scenarios, his estimations are about a population's imagined response to extraordinary cases. It neither measures real world responses to extraordinary cases, nor imagined responses to ordinary cases. The relationship between such experimental results and the ordinary use of concepts in the real world remains to be carefully drawn. However, as it stands, neither the experimental method nor the thought experiments used by Glasgow reveal how a term or concept is ordinarily used in the real world by a linguistic population.

Engaging with a thought experiment in an ordinary way should be understood as the outcome of an ordinary act of make-believe. Since the reality of race is up for debate in these cases, it is a safe assumption to look into the Reality Principle to flesh out something epistemically important. Unfortunately, the Reality Principle is not always an epistemically reliable guide for ordinary cases, because the ordinary concept of race is highly conventional and contextual. To imagine what is ordinarily believed about race, I cannot simply rely on my own views about the real world, not even by describing the ordinary communities of belief. I need to know what and how the ordinary person from a relevant community thinks about the real world, which is to say that I need to rely on the Mutual Belief Principle. If I, based on my individual beliefs, assumptions and commitments do not think or act in an ordinary way, then ordinary world-views need to supplement my interpretive tendencies. When seeking what passes for an ordinary case in make-believe, I need to bring to bear the assumptions and beliefs that ordinary people and relevant communities share regarding race.

6.1.2.1: The Tight Match. In the *Tight Match* thought experiment, I can easily imagine that OMB classification is simultaneously dominant with other racial classifications, popular or fringe. Personally, I can do more than just make-believe it, as I outright believe it and my life experiences corroborate it.

Racial terms are often used in vague or ambiguous ways. One is not always sure what classification is used or what if any, metaphysical theory is appealed to. This, I believe is an observable fact and a matter of life experience. All the same, prominent philosophers have addressed similar issues in their arguments. No less than Charles Mills provides seven different racial criteria which could come into conflict in such cases: bodily appearance, ancestry, self-awareness of ancestry, public awareness of ancestry, culture, experience, and self-identification (p. 50). Each criterion alone and in combination with others can justify some of the ways that racial language is ordinarily used. Most of the time, one may not be sure which instance of use relies on which criterion. One can try to be charitable in one’s interpretations and disambiguate along the way. A level of polysemy is generally expected, which fits in with the Mutual Belief

Principle's guarantee of tolerance and pluralism regarding background assumptions. No privileged hegemonic interpretation is required. Different communities of belief are justified to interpret thought experiments according to their views if they are meant to be participants. This means that even when the prevailing social situation disfavours a scientific classification, the Mutual Belief Principle might guarantee it.

6.1.3: The Authorized Game

6.1.3.1: George's Appearance-Transformation Machine. Some thought experiments admit a wider range of possibilities than anticipated. For Hardimon, this is why "Glasgow's thought experiments do not establish their intended result" (2017, p. 48). Even when participants play by the rules and follow along closely, they can generate contrary imaginings. For Hardimon, George's case "does not so much establish a result about race's content as raises a question about how the concept would be projected—if indeed it would be projected—in a world for which it was not built" (p. 47). Just as concepts shape expectations in the real world, they also shape expectations in make-believe. Simply put, projection is always conceptually dependent.

As stated earlier, some thought experiments admit a larger diversity of authorized imaginings than their authors intend. George's thought experiment is such a case. Based on all the salient facts known about the real world, participants are led to imagine that George belongs to a race. One of these facts is George's initial race, which Glasgow kindly makes fictional in the thought experiment. With it, participants can generate their imaginings based on the Reality Principle by concluding that what is transformed is mere appearance. In some ways, George's case is no different than the case of someone with extreme body modifications or even the performers of the Blue Man Group, neither of which is appropriately described as a case of racial transformation, without additional justification. There is a fact of the matter about race in these cases whichever way the concept of race is cashed out. If enough of the facts are dismissed, then the thought experiment would lean on the Mutual Belief Principle to consider what Glasgow and the ordinary members of the American English-speaking community would take to be salient. If so, participants would have to forget about George's pre-transformation race. Glasgow could have kept George's race a mystery, but he chose not to. This makes it difficult to say what anyone would imagine if they were not aware of this information.

What would George's race be post-transformation? Here, any option is authorized. Glasgow's stipulations are, sadly, not enough to control this thought experiment to serve his purposes. Due to this inadequacy, the results make no epistemic difference to Glasgow's argument. One learns more about Glasgow, if anything, than about the ordinary concept of race.

George's Appearance-Transformation Machine exhibits a clear problem with authorized games in thought experiments. As Hardimon has noted, there may be "no fact of the matter" settling George's post-metamorphosis racial status (2017, pp. 46-47). Participants are free to make-believe whatever they please. For instance, I could imagine that he is racially ambiguous, passing, or just 'different' after the transformation and, so far as I do not know enough about him, I could also imagine that physical features associated with race are the only ones transformed; that he has altered the perceivable bodily markers of racial membership, rather than altering his racial membership itself. Every option is authorized because there are insufficient and competing Principles of Generation at play.

As noted, Glasgow explicitly stipulates an “all-black ancestry” for George. But what does this mean exactly? The first problem here is vagueness. How far back does this claim about ancestry go? How many generations are considered? All of George’s ancestors, since the dawn of humanity? Could anthropology support such a claim? What little I know of anthropological history makes me doubtful of this possibility. I fail to see any easy way to qualify this claim, which makes me suspect that participants are not entirely sure of what they are mandated to imagine about George’s ancestry.

The second problem with Glasgow’s stipulation is ambiguity. Does it entail that George appears to be visibly “black” and his ancestors are members of the same race? If so, appearance and ancestry are assumed to be independent. If appearance is not viewed as a central feature or a necessary condition for race, the participant can imagine George’s transformation to be a change of appearance and not of race. On the other hand, if “all-black ancestry” means that George and all of his ancestors simply happen to be visibly “black” in their appearance, the participant is free to imagine a change of race. Here, either the participant holds appearance as a central feature or a necessary condition of race, or the participant suspects that the thought experiment is leading them to guess without providing enough information regarding George’s race (such as the race of his ancestors, the biological nature of the transformation, the status of his official identity records, and so on). These different interpretations authorize incompatible game worlds and participants are free to make-believe by picking and choosing whichever Principles of Generation they please. As a consequence, the thought experiment becomes epistemically unhelpful to the metaphysical debate over the significance of appearance and ancestry. What it does accomplish is to show the imaginative tendencies of participants in response to science-fiction scenarios.

6.2: Issues with the Mechanics of Generation

6.2.1: Endogenous Inferences

6.2.1.1: The Utopian Babies. In the *Utopian Babies* case from Joshua Glasgow, participants are to imagine the "Reboot" of almost all cultures and social organizations, together with the disappearance of every possible sign of race for the sole surviving generation of human infants. One may be tempted to imagine the lives of these infants as they grow and build a new society. But that is not where the thought experiment leads. Glasgow is interested in the immediate aftermath of everyone else dying, to ask: could it be that the race of these babies "vanishes" as soon as the last person capable of using these concepts dies? Does the epistemic loss of 'race' imply its metaphysical loss?

Perhaps one should look into the Principles of Implication in this case. The erasure of all racially relevant information and culture is the erasure of all sources of belief formation. The death of all humanity, except the lucky babies, is the death of all believers and all powers of belief, at least until the infants develop their capacities. No beliefs and no believers leave no access to any community of belief. With this privation, participants of the thought experiment are unable to imagine that the babies "have their races" as they supposedly did before the community perished. They are barred from the relevant beliefs and concepts, which means the use of the Mutual Belief Principle is deactivated in this thought experiment.

Participants are left with the Reality Principle. Based on their own beliefs about the actual world, could it be the case that in reality "the babies would still have their races after [everyone else] perishes, if they have any races to begin with"? Glasgow takes it to be so. He wants to make the point that one may believe humans have races and that races are ultimately identifiable independent of particular histories and practices. If participants believe that babies 'have' races in the actual world, this should imply that the babies in this thought experiment must 'have' races too. Nothing points to the contrary. When these imagined babies are taken to "have their races" under whatever racial classification, as long as this classification scheme is taken to be successful, they will continue to "have their races" after everyone else dies. They will "have their races" regardless of anyone believing or using the relevant racial classification because that is what it means to "have" something like race in the real world.

One way to understand this continuity is with scientific hindsight and foresight. For instance, even if before the Copernican revolution the astronomy of the day was geocentric, it does not mean that the sun was rotating around the Earth at the time. Beliefs about the real world do not easily change reality. Even when the geocentric belief is acknowledged and the sun orbiting the Earth is imagined, one's belief in the astronomy of the real world is not unsettled. On another note, if tomorrow the majority of people adopted the flat-Earth model, the rest of humanity would continue to believe in the more sensible theory as they should. Analogously, even when one imagines that all histories and practices of racial classification have disappeared, one is to believe that in the real world, race has not disappeared. For the fictional world to match the real world according to the Reality Principle, participants need to believe that the infants in question continue to "have their races" as before. If in one's imaginings, any of these babies were to 'have' a particular race, one continues with the same imagining regardless of whatever else is to be imagined. Unless one cannot imagine a baby to ‘have’ a race (whatever that means) one must continue to imagine the infant as before. It seems as if the Reality Principle supports this conclusion—does it not?

So far, I have overlooked something: these infants should not be said to “have their races” on their own. What exactly is it that these babies are supposed to possess? Glasgow’s answer is the “Asian-ness” of “a (racially) Asian baby” without further explanation (2019b, p. 133). This is baffling. If I had to have a wild guess (and I would rather not), I would think that he is implying that there are certain racial qualities or properties (‘features’ henceforth) that a baby has by virtue of which that baby is said to ‘have’ a race according to a certain racial classification. This would mean that whatever these features are that babies possess are different than the race they are associated with, as in whatever race a baby is said to ‘have’ cannot be the same features that the baby has independently of so saying. To make a further conjecture, I assume that these supposed features are unified under a particular race once enough of them fit within that race as a cluster (i.e., HPC Kind), such that enough “Asian-ness” is identified as “Asian” for Glasgow. It should be clear at this point that, even with my guesswork, I cannot make sense of what is exactly being implied. To think in terms of the Reality Principle means imagining based on my beliefs about the real world. But I don’t know what “Asian-ness” means in the real world other than a confused way of naming physical features commonly found within the “Asian” race. Still, this seems backwards. The babies are supposed to have their races by virtue of their features, rather than having these features by virtue of their race. The features are claimed to ground the race, not the other way around. But race grounding features are exactly how Glasgow frames it when he insists that the babies "have their races" in his thought experiment.

Glasgow's language suggests that the way these babies are in themselves and what is or could be said about them are two different matters. The participant could describe a baby they are imagining with a racial label. But that does not say anything specific about the way they imagine that baby. An "Asian" infant could be imagined with a variety of physical features found within the Asian population, such that two participants can describe dissimilar babies as "Asian" and refuse the same label for the description of each other's imaginings. This only means that racial guesswork is very much a possibility in this game world, just as it often is in the real world. The difference is that no confirmation is possible, because all racial information is erased. With Glasgow's questionable terminology, "Asian-ness" and being "(racially) Asian" are not the same thing, and the latter is based on the former. The babies cannot be said to "have their races" if what is meant by "their races" is something other than their features since no baby 'has' a racial conception or identity and everyone else is dead. These babies, therefore, could not be said to "have their races" if this 'having' rests on having a shared classification scheme.

Glasgow is trying to show that getting rid of the social determinants of race will not impact its non-social determinants. This is why he is going for a blank slate that only leaves infants behind. Babies raised by futuristic technology could not be said to take part in socially significant practices, because infants are not socialized humans. His strategy is as follows: if the infants could be said to "have their races," then it must be admitted that society is sufficient but not necessary for there to be races. The problem is that it is hard to get others on board with the claim that babies “have their races” if “races” are strictly about racial classifications, because classifications are not the kinds of things babies have. Supposedly, they have birth mothers too. But if the mothers die along with everyone else, I should not imagine the babies as having birth mothers after the fact, since they no longer ‘have’ birth mothers. Not everything that one could have, one could have independently and unconditionally.

The only one in the thought experiment with the relevant information is the participant. Effectively, the thought experiment asks whether these babies continue to “have their races” after everyone else dies by indirectly mandating participants to play the role of an adult with the racial baggage that was meant to be eradicated. Participants add back into their imaginings what the utopian plan erased. This tacit principle goes against the direct stipulations of the narrative, prematurely undermining the aim of the reboot. Glasgow does not explore the ways that the plan fails to achieve a racial utopia in the long run. As such, the utopian story becomes nothing but a distraction. It is no different than imagining several infants abandoned at a hospital without identification. In such a case, the participant could be mandated to imagine how even anonymous infants are said to “have their races” apart from the relevant social factors that usually inform racial identity. If everyone has a race, and these babies are included in everyone, then they too must “have their races” whether or not they have identification. This achieves the same results as the *Utopian Babies* thought experiment. Glasgow explains his strategy as “conceptual bootstrapping” and admits that it “may seem odd” but thinks it is useful nonetheless (2019b, p. 134). But this “bootstrapping” is nothing like the idea of rebuilding a ship that is sailing or ascending a ladder only to kick it away, nor does it have anything to do with its more technical uses in cognitive science and psychology. In speaking of “bootstrapping,” Glasgow is simply admitting that his move in this thought experiment raises suspicions that he is begging the question—the babies have their races because they are imagined to have their races.

6.2.1.2: The Disaster. The *Disaster* thought experiment summarizes the science fiction plot of the *Utopian Babies*. Glasgow argues that as soon “as the practice of racial classification stops,” all the infants stop having races (2009, p. 121). While the thought experiment is minimalist, capturing race under ‘practices of classification’ allows him to interpret anything associated with race as a racial practice. Indeed, thinking and acting are both loosely captured by ‘practices’ in this context, but they need not be. Strictly speaking, ‘practices of racial classification’ are all those practices that are involved in classifying races and nothing more. Such practices have shaped the lives of these children and the cultural resources they will inherit from the previous generations. So, even if they will not be practicing any racial classification as infants, once they are enculturated and socialized, they can come to practice racial classification. At the very least, the history and the cultural record of humanity will inform them enough to leave open that possibility. This obvious flaw in the *Disaster* thought experiment is amended in the *Utopian Babies* thought experiment by the destruction of historical records.

6.2.1.3: Temporary Amnesia. The *Temporary Amnesia* thought experiment, much like the *Utopian Babies* case, asks participants to imagine the end of racial thinking. Glasgow tries to show how race remains to be metaphysically real even when its epistemic or conceptual apparatus goes missing by imagining a momentary forgetfulness, rather than a wholesale cultural erasure. The difference is that this case is not open to racial practices in general, but involves only practices of racial classification and nothing more. It asks: if people lost their classifications for an hour, how could they apply them in practice? The problem is that social practices are not simply those that are consciously accessible. Instead of a *Temporary Amnesia*, Glasgow might as well try to show that, because race thinking and its social practices disappear when everyone goes to sleep, the salience of race is at an all-time low at night. But that would be a laughable suggestion. No one needs to be consciously aware of race thinking or practice.

The safer assumption would be that (almost) everyone lives in societies that are organized racially and their non-racially motivated pursuits can inadvertently maintain and reproduce the social regimes which configure into race. For example, I could easily fail to observe the racial division of labour present in the North American labour force, but my lack of evidence would not count as evidence against this practice. Whether I observe it or not, a racial division of labour continues to be a form of racial classification in practice. Such practices are historical and productive, and racial divisions produce racial effects which could be observed in later generations, even when those boundaries are suddenly abolished. These effects could be as simple as the practices of racial endogamy observed by tracing lines of descent, racial patterns of phenotypic distribution, and racially distributed cultural inheritance. In the end, racial practice precariously borders racial belief and racial phenomena and separating theory from practice can be difficult. What is not difficult is noticing how theory and practice do not always overlap. If the direct mandate is to imagine that everyone's "cognitive apparatuses short-circuit," the Reality Principle still indirectly mandates the imagining of ongoing background practices that reinforce racial patterns, because certain racial patterns are not mind-dependent. So, it should not be assumed that social practices of racial classification always depend on active mental states about race.

6.2.1.4: The Tight Match. The *Tight Match* thought experiment from Jeffers relies on a false dichotomy between social and natural metaphysical foundations. Consider, as an example, the standardized systems of units and measures. These systems, such as the Metric or Imperial systems of units, are 'fundamentally' social because they are the historical achievement of several social forces (cultural, political, legal, economic and so on). Such a system has to define a specific scale to uniformly apply and uniformly measure exact quantities. To best facilitate this latter requirement, systems of units are today fixed by specific 'standards' which are either prototypical physical objects or physical constants. A 'standard' is a reference that fixes a unit's extension and allows for calibration. However, a fixed standard is itself 'fundamentally' physical and not 'fundamentally' social, so it can be a neutral and non-arbitrary reference point. Standard systems of units are 'fundamentally' social institutions, but they are fixed by 'fundamentally' physical constants. Therefore, they exemplify a "tight match" between a cultural and a scientific classification, as Jeffers would have it. So, if such a system fails to fix its units of extension, then, following Jeffers, dominant ideas of the day should reappropriate its scale to demonstrate that the physical constant was inessential to the system all along and, inevitably, reorganize its standard into something 'fundamentally' social.

In 1834, both houses of the British parliament burned down. As it happened, they also housed the British standards of units. At the time, the British Imperial units were fixed by prototype physical objects. A brass rod prototype, built in 1760, became the definition of the ‘Yard’ by law in 1825 (1824, pp. 637-638). But the fire of 1834 irreversibly deformed the brass rod (1841, pp. 5-6). As a replacement, a prototype was set as the standard twenty-six years later (1855, pp. 589-592). This means that between 1834 and 1855, the British Imperial system had no legal standard of measurement for what the length of a ‘Yard’ should be. So, in following Jeffers, I ask: did the absence of a prototype for twenty-six years show that standardization was inessential to the system, and did this result in an inevitable reorganization of that system based on dominant ideas of the day? The system of measurements remained unchanged even without a fixed standard, not because a prototype was not needed, but because there were enough reference copies of the prototype available to allow approximate calibration, even if with a slightly higher tolerance for error. Thus, between 1834 and 1855, the “tight match” between the system of units and its standard became ever so slightly loosened, but a misalignment did not occur.

Systems of units are both social and physical. At the same time, they are the paradigm of what it means to be ‘objective’ in the sciences. Just as these standardized systems have both social and physical foundations, a racial classification, as a system of human variation, could be both social and physical. The “tight match” imagined in the thought experiment is such a system. To align common sense terminology to the scientific system in question is to closely fix racial terms with their scientific referents in population genetics. It is worth emphasizing that Spencer does not advocate for such a project, and neither do I. But the *Tight Match* thought experiment does indulge this fictional possibility. What the Yardstick example shows is that a subsequent shift away from the “tight match” need not support what Jeffers confidently argues that it does;

namely, that “a reorganization of racial designations and identifications” would necessarily incur and that a “previous connection to something more biologically significant would thus be revealed as inessential.” Just as in the Yardstick case, there may be enough physical and biological reference points entrenched in society, culture and popular ideas to allow for the approximation of physical or biological foundations, even if with a slightly higher tolerance for error.

6.2.1.5: The English-Bangladeshi Woman. Consider a simple question about the *English-Bangladeshi Woman* thought experiment: why should anyone be concerned with what the young woman of the thought experiment thinks? Perhaps, because Jeffers wants participants to see that informing common sense racial classifications with scientific ones confuses the core notion of race that allows physical appearances to be associated with ancestry. Except, in the thought experiment, the young woman is said to be aware of her Bangladeshi ancestry. So, it is unclear why she should find the scientific classification at all confusing. The classification Jeffers uses identifies her as a Bangladeshi (or English-Bangladeshi) “Caucasian” or “Caucasoid” (Jeffers, 2019a, p. 43). These labels could not possibly confuse her belief in her Bangladeshi ancestry, because she knows herself to be, and would continue to be, born of (supposedly Bangali) parents from Bangladesh. There is an incongruency between Jeffers’ worry about the possible confusion caused by the reform of common sense, on the one hand, and the fictional truths of the thought experiment, on the other.

Suppose that I felt a worry about the possible confusion that could ensue from the correction of the young woman’s common sense. If so, I would have to think that when she reforms her conception of race, the way she articulates her racial predicament becomes confused. How? Supposedly because “it conflicts with common sense” (Jeffers, 2019a, p. 44). Whose common sense? It could not be hers, because reforming one’s common sense beliefs either results in reformed common sense beliefs or in additional beliefs. If her beliefs are reformed, in line with Jeffers’ wording, she could say “My darker skin has marked me for my whole life as a minority of foreign origin and my appearance has generated a particular racial experience for me, even though I am part of the same broad racial family as the majority population in Britain.” She could also summarize her point and say “My appearance in this country is racial.” Conversely, if she gained additional beliefs, she could say “My darker skin has marked me for my whole life as a minority of foreign origin and my appearance has generated a particular racial experience for me, even though technically I am part of the same broad racial family as the majority population in Britain.” Again, she could summarize her point and say “My appearance in this country is racial.” There simply is no practical difference between these two cases and no confusion has to ensue. Not only could she articulate her racial experience very well, but there is no good reason to imagine that she forgets her unreformed position. If the context demanded, she could revert to her older ways of articulating herself. This is expected since people often alternate between ways of thinking and talking depending on context. Here, the argument falls short of the “conflicts with common sense” that it promised.

I retrace some of the basic moves of this thought experiment. Someone is imagined to hold common sense beliefs. These beliefs are then reformed. She subsequently has reformed beliefs. Now, is it a good idea for her to hold these reformed beliefs? The answer is assumed to be negative because her reformed beliefs come in conflict with common sense beliefs. If this is correct, it feels as though the 'conflict' in question is just a pejorative term for 'reform' and the thought experiment is inexplicably biased in favour of common sense beliefs.

6.2.2: Exogenous Inferences

6.2.2.1: The Tight Match. By way of the *Tight Match* thought experiment, Jeffers argues that, metaphysically speaking, "race is fundamentally social and not fundamentally biological" and that people's "looks and lineages" tie them in essence to a place of origin (2019b, pp. 182-183). This seems like a contradiction if he is tracking the same ancestral lineages that population genetics tracks. But, for him, one's "lineage" is not independent of one's "looks" since, I take it that, they work importantly together. As he states, the "forms of physical difference involved in racial distinctions are necessarily at least partially related to forms of reproductive isolation, whether as a result of people being geographically separated going back to the distant past or through more recent social distinctions" (p. 182). Thus, any biological pattern associated with race is thought to be explainable by cultural practices going back generations. For example, if contemporary populations can be statistically clustered into continental populations, it is because today's populations have maintained their "lineage" through their particular cultural practices, such as marriage and child-rearing.

The *Tight Match* thought experiment is motivated, at least in part, by the claim that Spencer "wrongfully downplays the independence" of biological and social aspects of race (Jeffers, 2019b, pp. 181-182). This is an interesting point and one that deserves attention. However, I find it to be misdirected because Spencer is openly and emphatically defending a "radically pluralist" position (2019a, pp. 78, 82; 2019b, 211, 213, 238). Spencer defines his radical pluralism as one that does not presuppose "a single dominantly correct answer to the question of what race is and whether it's real in the relevant context, but [presupposes there to be] at least one dominantly correct answer to this question" (p. 213). This means that there could be multiple correct answers, while none of the correct answers are currently dominant. In contrast to the conclusion that "race is fundamentally social and not fundamentally biological," radical pluralism rejects the very notion of fundamentality. There are different ways of talking about race, one dominantly correct form of which, in this context, serves as a proxy for the biological model where others may not. Every little misalignment of the OMB classes with the groups from population genetics is accounted for by this pragmatic approximation. It is not perfect by any means, but useful enough to do the job Spencer has in mind. Other inquiries should either be directed at the population genetics study or the OMB classification itself, neither of which is understood to be metaphysically fundamental. If anything, the OMB classification is a (fundamentally) social classification that serves as a proxy for the (fundamentally) biological "K = 5" classification. Spencer is simply showing how the two classifications are in close alignment. Talk of fundamentality is, thus, redundant.

Once Spencer's radically pluralist theory is accepted, imagining "what is essential to race" becomes irrelevant. Participants are free to play the game of make-believe that Jeffers has contrived, but neither the participants nor Jeffers could decisively justify a conclusion for or against Spencer's position based on the thought experiment. Informing participants of the details of Spencer's theory is not always like explicating what remains implicit within the narrative of the thought experiment. Even with close attention to the

Principles of Generation that Jeffers has stipulated, they would fall short of learning about Spencer's radical pluralism. Instead, they are likely to follow Jeffers and imagine that Spencer "wrongfully downplays the independence" of biological and social aspects of race, and wrongfully presume him to be concerned with essence and metaphysical fundamentality. Jeffers is careful in his language not to directly make such a claim about Spencer. Pressed on the issue, I suspect, he would clarify that he is only targeting a hypothetical position, one that holds biology to be the metaphysical foundation of race. However, his narrative is ripe for misinterpretation. Spencer is not even risking such a hypothetical position, but the naïve participant could easily interpret the prescribed "tight match" to be Spencer's endorsement—a strawman position.

The *Tight Match* is misleading because Spencer's position is not properly presented. At least, clarifications should be added before or after the thought experiment, because learning about such contextual information is beyond the regular responsibilities of participants. The Mutual Belief Principle only guides participants to interpret the fictional world through the perspective of the work's author, the community of belief the author belongs to, or the beliefs of the work's intended audience. It has nothing to do with the competing beliefs or the competing community of belief challenged in the argument, which leaves the responsibility of providing the context squarely on the author.

6.2.2.2: The English-Bangladeshi Woman. The *English-Bangladeshi Woman* thought experiment mandates participants to imagine being the young woman of the narrative, or at least imagine what it would be like experiencing life in her social position. This is because of an indirect Principle of Generation mandating participants to imagine *de se* what the young woman “should” think. However, this mandate gives rise to an exogenous problem.

As my exposition of the metaphysics of race debate should have made clear, it is not always plainly evident what the language in the debate is assigned to capture, nor is the scope of each argument pronounced in detail. For instance, it is not always explicitly clear that this philosophical debate is not about all of the different ways people have divided human races historically and in different societies. Generally speaking, in philosophy, parochialism tends to be doctrinal rather than geographical, and commitments tend to be theoretical rather than cultural. So, it could come as a surprise that a debate over metaphysics is actually about the mostly ordinary Anglo-American uses of the term ‘race’ and mostly about the racial history and politics of the United States.

Consider, as an example, my peculiar failure to discover in time that the debate I took myself to be studying is overwhelmingly a debate I am excluded from the outset. I, as an Iranian-Canadian migrant and a second-language speaker of Canadian English, may take myself to be within the fuzzy borders of the American English-speaking community. But for this thought experiment, my position is extraneous for several reasons. At the very least, I am not anything remotely similar to a representative of the American English-speaking community, even when considering the dominant trends in this population. This is because, according to Jeffers, I do not share this linguistic community’s ‘common sense’ responses.

The *English-Bangladeshi Woman* thought experiment can be seen as a mismatch between the Reality Principle (i.e., the scientific classification) and the Mutual Belief

Principle (i.e., the common sense classification). As one would expect, such a mismatch may be remedied by the scientific classification correcting common sense misconceptions, which is to say the Reality Principle overrides the Mutual Belief Principle. But, for Jeffers, abandoning the "core notion" of race inevitably leads to an unappealing "confusion" for the self-understanding of the common person (2019a, p. 43). To prevent this, he recommends preserving common sense from scientific reformation, which is to say defending the Mutual Belief Principle from the Reality Principle. Yet, as Spencer points out, claiming that common sense is "more illuminating" rests on a purely "intuition-based" move (2019b, p. 233). This is because common sense beliefs and common sense responses are not the same. Participants cannot generate common sense responses even when supplementing with the right common sense beliefs, because 'common sense responses' are another way of talking about 'intuitions.' The best thing to do is be cognizant of the shared 'common sense responses' of the relevant community in question.

Amongst the American English-speaking community's distributed members, I am an outlier and my sensibilities are uncommon. Neither my personal life experience nor my membership in a minority social stratum in the Anglo-Canadian context is relevant unless shaped by this larger linguistic community's dominant patterns of common sense reaction. If members of this vast community were clustered into cultural groups to represent dominant trends, it would still not help the situation since one of the most important pillars of Jeffers' cultural constructionism is "participation in distinctive ways of life" (2019a, p. 50). Being Canadian, or a Canadian immigrant, makes me a participant of a distinctive (enough?) way of life, but still leaves me an outlier of the North American

English-speaking community at large. This being said, Jeffers is not even interested in the entire community. He wants participants to take into account what the young English-Bangladeshi woman should make of the scientific conception. The trouble with this mandate is that the Mutual Belief Principle requires participants to *de se* imagine her 'common sense' response. However, for Jeffers, "racial consciousness" is itself cultural, the means of cultural self-propulsion, and the by-product of a shared racial life (2019a, p. 64). Subsequently, participants not sharing in this racial consciousness should not be authorized to partake in *de se* imagining from the inside because they are barred from appealing to the Mutual Belief Principle. To use a crude example, "racial consciousness" is analogous to "pregnancy consciousness" because its share is not equally distributed. I cannot, nor should I dare to, imagine what someone pregnant should make of a scientific classification that revises their common sense notions. The best that I could do is to take account of the pattern of response I observe from those with the appropriate consciousness and adhere to the Reality Principle to supplement real world observations for my *de se* imaginings. This may not be what Jeffers anticipated but, according to his theory, it is the best that participants who do not share the appropriate consciousness can do.

Counter to my Canadian identity, it seems possible to argue that, as an Iranian, I participate in a distinctive way of life akin to what Jeffers has in mind. Although I do not view the identity 'Iranian' to be best described as a race, my Iranian 'way of life' satisfies the key aspects of cultural significance for the social construction of race that Jeffers specifies: being Iranian as cultural; as facilitating new cultural developments; and as shaped by prior cultural developments (Jeffers, 2019a, pp. 64-65). In addition, being

Iranian satisfies the logical core of the concept of race: it involves relevant visible physical features; it is linked by common ancestry; and it originates from a distinctive geographical location (pp. 39-40). Thus, being Iranian is what Jeffers would call "fundamentally sociohistorical" (pp. 38-50). Unfortunately, this is irrelevant to Jeffers' project. "Races are," for him, "appearance-based groups that initially result from the history of Europe's imperial encounters" (p. 65). Being Iranian is not the appearance-based product of the European age of empires, since the European encounter and recognition[40] of Iranian peoples goes back to, and has been continuous with, the dawn of anything European. Thus, being Iranian cannot be a race according to this "Imperial encounters" condition, because Iranians are racially classified earlier than this historical period. So, even if English-speaking Iranians, Iranian-Canadians or Iranian-Americans were to be clustered together in the American English-speaking community, the distribution of their common sense responses would be irrelevant. Simply put, such cultural communities do not qualify as races and their common sense does not stem from racial consciousness, even if they believe otherwise. They may even misunderstand what to do about other races when it comes to the Mutual Belief Principle because whatever they count as a shared racial way of life is not racial according to Jeffers.

In the *English-Bangladeshi Woman* thought experiment, participants are mandated to imagine what the young woman should think, while Jeffers defends a model

[40] Even the philosophical works of the Hellenic and Roman periods bear mention of the 'Persians' (coined as the name for most Iranians since the Achaemenid Empire). Leaving aside what the ancients thought about race, the racial connotation of 'Persian' or 'Iranian' in premodern Europe was commonplace enough that the chronicler Robert the Monk's retelling of a speech by Pope Urban II talks of a "wicked race" that is "a race from the kingdom of the Persians, an accursed race, a race utterly alienated from God, a generation forsooth which has not directed its heart and has not entrusted its spirit to God" ([1897]1971, p. 5). Disregarding how it misidentifies the frontiers of the Iranian civilization in an attempt to refer to the Seljuk Turks, its racially divisive language in rousing support for the First Crusades—an event *prior* to the European age of empires—is unmistakable.

of cultural construction that makes race a roughly homeostatic cultural community striving to maintain a stable but exclusive way of life. This model is, for many participants, incompatible with the mandate of the thought experiment, given at least two commitments: first, the three key aspects of cultural significance for the social construction of race articulated in the language of "racial consciousness" (Jeffers, pp. 64-65); and, second, the normative commitment that "we," understood as a racial population's internal and exclusive pronoun,[41] "ought to actively continue constructing races as cultural groups" (p. 58). Once participants accept the boundaries of the debate and adhere to a Jeffers-style conception of race, they still cannot all be authorized to substitute the young woman's response if they are not internally connected to the sociohistorical and cultural community to which she belongs.

Added precision helps to prevent over-generalization in arguments, but it can often complicate thought experiments. In the metaphysics of race debate, 'common sense' is a more expansive qualifier, extending over both 'manifest' and 'operative' concepts, whereas 'ordinary' is associated with the 'operative' concept alone. Here, 'manifest' captures what someone takes themselves to be applying or attempting to apply in some cases, while 'operative' captures what someone does in practice (Haslanger & Saul, 2006, p. 99). Jeffers is after both because he is concerned with common sense. The problem is that without appropriate 'racial consciousness' access to the manifest concept

[41] My qualification of the pronoun may be challenged. Yet, I find no other sensible way to interpret the passage, given how an inclusive pronoun would be an open invitation to all those activities or processes which have constructed race since the European imperial encounters. But they are not all commendable. Even if the "ought" within this passage was limited to the epistemic construction of race, it would be an ought too far. For Jeffers, racial identities are self-understandings afforded by racial consciousness. This is internal and exclusive through and through. Unless the normative weight of this "ought" is watered down enough to only allow the preservation of what the reader deems worthy of conservation or construction—an incredible project to be sure—it must be admitted that, at the very least, there cannot be a passive moral imperative.

is denied. What remains is the operative concept which is associated with the ordinary concept, but not what Jeffers is after.[42]

I move on to Jeffers' historical account of racial genesis, which I suspect, inadvertently commits him to a mild Euro-centric position. I am not underestimating the significance of this history, and I doubt that anyone could understand the contemporary phenomenon of race without it. At the same time, I do not overestimate its explanatory power. For instance, I can articulate at least one credible way that the conclusion drawn from the *English-Bangladeshi Woman* thought experiment is unconvincing because of this historical commitment: the young woman in question does not even share in the common sense 'intuitions' of the American English-speaking community that Jeffers belongs to and is concerned with. Similar to my own case, the English-Bangladeshi woman is atypically situated along the fuzzy margins of this vast linguistic population and her response does not represent the "consciousness" that Jeffers associates with race. As such, the appropriate community of belief for *de se* imagining the English-Bangladeshi woman is not accessible to Jeffers. Presumably, the woman's parents are from Bangladesh, which makes her the child of recent migrants to Britain. Therefore, her racial experience is also captured by the recent history of global migration patterns, not merely the history of Europe's imperialism. As it turns out, focusing on the young woman's possible response has distracted from the fact that an English-Bangladeshi case does not authorize Jeffers to *de se* imagine a response. Given his conception of race, the

[42] What has been called the "target" concept—what one should, ideally, be employing—does not apply in this context (Haslanger & Saul, 2006, p. 99). Here, *de se* imaginings cannot be generated according to target concepts because 'what one should ideally employ' in a thought experiment is just what is mandated by that thought experiment to be imagined as fictional truth. Principles of Generation are at work, and aligning a participant's practices and intentions makes no sense in this context.

Mutual Belief Principle cannot help his *de se* imagining of her since he does not share in *her* racial consciousness and common sense.

The imagining of the young woman's response is a move authorized in the thought experiment, but it remains a move unauthorized for someone like Jeffers given his own theoretical commitments. Jeffers' position implies that mutuality/solidarity between races should not be confused with mutuality/solidarity within races. The latter, which he calls "racial consciousness," should be the guide, but it is the former that compels him to conclude in this case. This is an error according to his own theory. Appearance-based imaginings do not make-believe racial consciousness, and imagining racial consciousness does not make-believe based on appearances. It should now be clear why committing to a genesis of "appearance-based groups that initially result from the history of Europe's imperial encounters" is more Anglo-American a history than anticipated. It is far too exclusive for the entire American English-speaking community and, for the lack of a better word, it is far too provincial to accommodate Jeffers' own conception of race. I think this exogenous problem is due to the concern Jeffers shares with Haslanger. Their focus on racial hierarchies of subordination has left a blind-spot for racial systems of exclusion.[43] In this case, this is true to such a degree that Jeffers has inadvertently designed a thought experiment that excludes himself from participation.

[43] I borrow this difference from George M. Fredrickson's distinction between the racism of inclusion and the racism of exclusion: "the inclusionary variant permits incorporation only on the basis of a rigid hierarchy justified by a belief in permanent, unbridgeable differences between the associated groups, while the exclusionary type goes further and finds no way at all that the groups can coexist in the same society" (2002, p. 9).

Chapter 7: Conclusion

In this thesis, I have argued that thought experimentation is, as a method, reliant on the power of imagination. The misapplication of thought experiments can be diagnosed through the lens of Walton's theory of make-believe when thought experiments are taken to be epistemically valuable imaginings generated in response to fictional narratives. Thought experiments go wrong when their fictional content does not achieve its planned epistemic aims, either because of the Machinery of Generation (i.e., the Reality Principle, the Mutual Belief Principle or the authorized game), or the Mechanics of Generation (i.e., endogenous or exogenous inferences) at play. The theory of make-believe conveniently allows for the isolation of methodological problems in thought experimentation, without metaphysically loaded explanations.

In Chapter Five, I provided an exegesis of Walton's theory of make-believe, with added revisions to suit the methodological needs of thought experiments. I discussed how thought experiments are to be understood as fiction. I clarified Walton's view of 'mimesis' through Gombrich's substitution model. I made the case that the theory of make-believe is best seen as an enhanced and expanded version of reader-response theory. I argued for a careful way of understanding 'make-believe' that seeks to avoid unnecessary phenomenological, epistemic or ontological confusion. Then, I detailed Walton's Machinery and Mechanics of Generation, followed by some revisions tailored for the methodological study of thought experiments.

The thought experiments I use as my case studies are sourced from the literature on the metaphysics of race, particularly from the works of Joshua Glasgow (the *Twin-Earth*, *George's Appearance-Transformation Machine*, the *Dalai Lama*, the *Utopian*

Babies, the *Disaster* and the *Temporary Amnesia*) and Chike Jeffers (the *Tight Match* and the *English-Bangladeshi Woman*). By subjecting each case to the scrutiny of the theory of make-believe, I have been able to pinpoint the cause of their controversial reception. In total, I detected at least twelve instances of misapplication (three cases involving the Reality Principle; one involving the Mutual Belief Principle; one unauthorized game; five endogenous problems of inference; and two exogenous problems of inference). I believe the criticisms I have made of these examples easily demonstrate the analytical power of the theory of make-believe in resolving methodological confusion over thought experiments.

My motivation for focusing on the selection of case studies in this thesis is two-fold: firstly, there is an open methodological disagreement over the use of thought experiments in the metaphysics of race which needs to be addressed; and, secondly, the particular thought experiments selected are not only suitable because they conveniently exhibit the power of the theory of make-believe, but because they are sourced from a current debate between some of the best theorists in the field.

Theorizing about race is highly controversial, often for good reason. Yet, this frequent controversy can itself be controversial if it gets in the way of sensible and justified inquiry into topics associated with race. This is why I entirely dedicated Chapter Three and Chapter Four to explain what is at stake in the metaphysics of race debate so considered. In Chapter Three, I catalogued the central questions, the terminology, and the concepts needed to understand what is metaphysically at stake in this debate. Then, in Chapter Four, I used the working definitions I developed in Chapter Three to meticulously compare the differences between the positions engaged in the debate. I believe that these two chapters should clear up any reasonable worries about the risk of misunderstanding in my thesis.

7.1: Further Research Questions

7.1.1: Implication and Inference

In Chapter Five, I claimed that ‘Principles of Implication’ are best understood as ‘Principles of Inference’ because they guide participants to imagine indirect fictional truths as a consequence of prompts or direct Principles of Generation. My reason was that implication is, and inference is not, a relation between statements. Walton captures fictional statements under ‘fictional proposition’ and insists that this unit does not carry the same theoretical import as ‘proposition’ ordinarily does in philosophy. Based on this, I argued that there cannot literally be any relation between antecedent and consequent propositions when indirect fictional truths follow imaginings generated by prompts or direct fictional truths in make-believe.

The trouble is that Walton’s ‘fictional proposition’ could be mistaken for ‘fictive proposition’ which is a way of describing the propositions metaphysical fictionalism concerns itself with. Fictionalism is the view that the usefulness or acceptability of statements made within a domain of discourse can be explained by treating them as, or as similar to, fictional statements (Manley, 2015, p. 358). Fictionalists disagree about the intent of the speaker uttering a fictive statement, or about the status of fictive statements, but they agree that they are talking about the issue of empty reference in statements with propositional value. Fictive statements are modified by fictive sentential operators (a.k.a. “story prefix” (Rosen, 1990, p. 331)) such that their truth value is excused according to some truth in the world—usually the source material of the fiction—such that, for example, “it is true according to a work of fiction that so and so is the case” is considered a true sentence (Lewis, 1978, p. 37). Walton’s theory does not work with fictive operators, and whatever sentential operators Walton uses make no mention of their source material. “It is fictional that” is a reconstruction of a mandate, prescription or prompt that a participant responds to in a game of make-believe, rather than a reference to a work of fiction. Confusing Walton’s ‘fictional propositions’ with ‘fictive propositions’ could result in the view that since empty names make no compositional contribution to propositions and leave “gappy” truth values, many presumed fictional propositions express no propositions at all (Yablo, 2021, p. 347). One could only argue this if one confused the mandate of fictional truths for propositional attitudes, and took fictional propositions to be more than mere heuristic devices. Things can get even more confused if propositions, instead of abstract entities, are accounted for as epistemically useful fictional devices.

I have not expanded on how Principles of Inference are to be understood in contrast to Principles of Implication—if indeed they are to be understood differently. Because of this, I have not explained how what I call ‘tacit’ Principles of Generation are to be understood in relation to inferences. Instead, I have simply used tacit principles as if they are Principles of Inference to find endogenous problems in thought experiments. The problem is that I have defined tacit principles as dispositional and, in light of my discussion in Chapter Three, mind-independent. I cannot say what mind-independent inferences are meant to be, or whether inferences can be mind-independent. But I acknowledge that it needs further clarification.

7.1.2: The Problem of Intentionality

Issues around 'fictional propositions' are only suggestive of a deeper complication in Walton's theory. A keen observer may notice that Walton's move in collapsing the distinction between the so-called 'representational arts' and 'works of fiction' does more than risk a phenomenological ambiguity between imagination and perception. It can equivocate between the *object* which Walton calls a 'prop' or a 'prompter' (e.g., a text) and the *content* generated by prompts or direct Principles of Generation (e.g. a narrative). For instance, at some point, the theory moves from marked pieces of paper to words and sentences of a language as its props. The representational object (i.e., the extension) becomes confused with the object of representation (i.e., the intention) and, consequently, the object of intentionality becomes confused with the intentional object. Such confusion between a mental state, or its expression, and the thing it is about is sometimes called 'the problem of intentionality' (Blackburn, 2005, p. 188).

Consider Walton's favourite example, whereby a group of children playing in the woods agree that "all stumps are bears" which provokes them to automatically and unreflectively imagine a bear when they see a stump and be triggered to "jump with fright" (1990, p. 24). The stump is a prop, and "all stumps are bears" is a direct Principle of Generation. The problem is that something like "bears are frightening" must also be an indirect Principle of Generation (which I call a tacit principle) in the game. This mandates something about bears (i.e., the content) rather than stumps (i.e., the object). But, how are the content of imaginings directed at non-existent entities?[44]

[44] Adam Toon has raised this question about Walton's theory in the context of scientific modeling (2012, p. 81).

There is a possible solution to the problem of intentionality so posed, but I can only leave it as a suggestion here. In later writings, to explain metaphors, Walton differentiates between ‘content-oriented’ games and ‘prop-oriented’ games of make-believe: content-oriented games mandate fictional content or mandate fictional content in response to props, whereas prop-oriented games mandate fictional content about props (2015, p. 175). In his treatment of metaphors, Walton takes words to be capable of being *ad hoc* props (p. 178). This means that in prop-oriented games, fictional content can be treated as a prop. In the example above, stumps are the props that are to be imagined as bears, and the imagined bears are the fictional content in a content-oriented game. But the imagined bears are also to be imagined as scary by the children, which makes imagined bears props for imagining fright in a prop-oriented game. Therefore, in this example, the fictional bears are firstly fictional content to be imagined and secondly props for imagining fright.

Most games of make-believe involve a blend of both content-oriented and prop-oriented games, but this is not always apparent nor always the case. Museums and art galleries are full of actual artifacts that are props with which people play content-oriented games. In contrast, dictionaries and thesauruses are full of words with which people play prop-oriented games. Of course, the books called ‘Dictionary’ and ‘Thesaurus’ are objects, but the content-oriented games people play with them as books (e.g., “Something means so and so because the book says so, and the book says so because it means so and so” as circular reasoning; etc.) are different from the prop-oriented games people play with them (e.g., “Something means so and so because it frequently co-occurs with such and such in these linguistic contexts” as collocative or reflected meaning; etc.). Thus, it may not be difficult to sort particular imaginings into content or prop-oriented games. However, sorting complex games of make-believe into one or the other seems difficult and doing so could call for an overhaul of the Mechanics of Generation.

References

An Act for Ascertaining and Establishing Uniformity of Weights and Measures. (1824, June 17). London, United Kingdom. Retrieved August 2023, from https://www.legislation.gov.uk/ukpga/1824/74/pdfs/ukpga_18240074_en.pdf

An Act for Legalizing and Preserving the Restored Standards of Weight and Measures. (1855). In *A Collection of the Public General Statutes: 1855* (pp. 588-592). London: George Edward Eyre and William Spottiswoode.

Aristotle. (1963). *Aristotle's Categories and De Interpretatione.* (J. L. Ackrill, L. Judson, Eds., J. L. Ackrill, Trans.) London: (Clarendon Press) Oxford University Press.

Ayer, A. J. (2005). Russell, Bertrand Arthur William (1872–1970). In J. Rée, & J. O. Urmson (Eds.), *The Concise Encyclopedia of Western Philosophy* (pp. 336-342). New York: Routledge.

Bach, K. (2015). Type-Token Distinction. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (p. 1089). New York: Cambridge University Press.

Blackburn, S. (2005). *The Oxford Dictionary of Philosophy.* Oxford: Oxford University Press.

Bostrom, N. (2009). Ethical Issues in Advanced Artificial Intelligence. In S. Schneider (Ed.), *Science Fiction and Philosophy: From Time Travel to Superintelligence* (pp. 374-382). Malden, MA: (Blackwell Publishing) Wiley-Blackwell.

Bourdieu, P. ([1981] 2014). Men and Machine. In K. Knorr-Cetina, & A. V. Cicourel (Eds.), *Advances in Social Theory and Methodology: Toward an Integration of Micro- and Macro- Sociologies* (pp. 304-17). London: Routledge.

Bourdieu, P. (1990a). *The Logic of Practice.* (R. Nice, Trans.) Stanford, CA: Stanford University Press.

Bourdieu, P. (1990b). *In Other Words: Essays Towards a Reflexive Sociology.* Stanford, CA: Stanford University Press.

Bourdieu, P., & Wacquant, L. J. (1992). *An Invitation to Reflexive Sociology.* Chicago: University of Chicago Press.

Boyd, R. (1989). What Realism Implies and What It Does Not. *Dialectica*, *43*(1-2), 5-29.

Bromberger, S. (1992). *On What We Know We Don't Know: Explanation, Theory, Linguistics, and How Questions Shape Them.* Chicago: University of Chicago Press.

Butchvarov, P. (2015). Conceptualism. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (p. 194). New York: Cambridge University Press.

Butchvarov, P. (2015). Metaphysics. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (pp. 661-664). New York: Cambridge University Press.

Delaney, C. F. (2015). Hypostasis. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (p. 487). New York: Cambridge University Press.

Durkheim, E. (1982). *The Rules of Sociological Method and Selected Texts on Sociology and its Method.* (S. Lukes, Ed., & W. D. Halls, Trans.) London: (The Free Press) The Macmillan Press.

Felluga, D. F. (2015). Reader, Reading and Reader-Response Criticism. In D. F. Felluga, *Critical Theory: The Key Concepts* (pp. 260-264). New York: Routledge.

Fish, S. (1980). *Is There a Text in This Class? The Authority of Interpretive Communities.* Cambridge, MA: Harvard University Press.

Fish, S. E. (1976). How to do Things with Austin and Searle: Speech Act Theory and Literary Criticism. *MLN*, *91*(5), 983-1025.

Fredrickson, G. M. (2002). *Racism: A Short History.* Princeton, NJ: Princeton University Press.

Friend, S. (2008). Imagining Fact and Fiction. In K. Stock, & K. Thomson-Jones, *New Waves in Aesthetics* (pp. 150-169). London: Palgrave-Macmillan.

Gendler, T. S., & Liao, S. Y. (2016). The Problem of Imaginative Resistance. In J. Gibson, & N. Carroll (Eds.), *The Routledge Companion to Philosophy of Literature* (pp. 405-418). London: Routledge.

Glasgow, J. (2009). *A Theory of Race.* New York: Routledge.

Glasgow, J. (2019a). Glasgow's Reply to Haslanger, Jeffers, and Spencer. In J. Glasgow, S. Haslanger, C. Jeffers, & Q. Spencer, *What is Race?* (pp. 245-273). New York: Oxford University Press.

Glasgow, J. (2019b). Is Race an Illusion or a (Very) Basic Reality? In J. Glasgow, S. Haslanger, C. Jeffers, & Q. Spencer, *What is Race?* (pp. 111-149). New York: Oxford University Press.

Glasgow, J., Haslanger, S., Jeffers, C., & Spencer, Q. (2019). *What is Race? Four Philosophical Views.* New York: Oxford University Press.

Godfrey-Smith, P. (2003). Goodman's Problem and Scientific Methodology. *The Journal of Philosophy*, *100*(11), 573-590.

Gombrich, E. ([1963] 1978). Meditations on a Hobby Horse or the Roots of Artistic Form. In E. Gombrich, *Meditations on a Hobby Horse: and Other Essays on The Theory of Art* (pp. 1-11). London: Phaidon.

Goodman, N. (1955). *Fact, Fiction and Forecast.* Cambridge, MA: Harvard University Press.

Grenfell, M. (2014). Field Theory - Beyond Subjectivity and Objectivity: Introduction. In M. Grenfell (Ed.), *Pierre Bourdieu: Key Concepts* (pp. 41-47). New York: Routledge.

Grieve, G. (2001). Fetish. In V. E. Taylor, & C. E. Winquist, *Encyclopedia of Postmodernism* (pp. 120-121). New York: Routledge.

Guter, E. (2010). *Aesthetics A–Z.* Edinburgh: Edinburgh University Press.

Hacking, I. (1991). A Tradition of Natural Kinds. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *61*(1/2), 109-126.

Hacking, I. (2007). Natural Kinds: Rosy Dawn, Scholastic Twilight. *Royal Institute of Philosophy Supplement*, *61*, 203-239.

Hakli, R., & Mäkelä, P. (2015). Social Ontology. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (pp. 999-1000). New York: Cambridge University Press.

Hardimon, M. O. (2003). The Ordinary Concept of Race. *The Journal of Philosophy*, *100*(9), 437-455.

Hardimon, M. O. (2017). *Rethinking Race: The Case for Deflationary Realism.* Cambridge, MA: Harvard University Press.

Harper, W. (2015). Natural Kinds. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (p. 701). New York: Cambridge University Press.

Haslanger, S. (1995). Ontology and Social Construction. *Feminist Perspectives on Language, Knowledge, and Reality*, *23*(2), 95-125.

Haslanger, S. (2012a). A Social Constructionist Analysis of Race. In S. Haslanger, *Resisting Reality: Social Construction and Social Critique* (pp. 298-310). New York: Oxford University Press.

Haslanger, S. (2012b). What Are We Talking About? The Semantics and Politics of Social Kinds. In S. Haslanger, *Resisting Reality: Social Construction and Social Critique* (pp. 365-380). New York: Oxford University Press.

Haslanger, S. (2012c). What Good Are Our Intuitions? Philosophical Analysis and Social Kinds. In S. Haslanger, *Resisting Reality: Social Construction and Social Critique* (pp. 381-405). New York: Oxford University Press.

Haslanger, S. (2019a). Haslanger's Reply to Glasgow, Jeffers, and Spencer. In J. Glasgow, S. Haslanger, C. Jeffers, & Q. Spencer, *What is Race? Four Philosophical Views* (pp. 150-175). New York: Oxford University Press.

Haslanger, S. (2019b). Tracing the Sociopolitical Reality of Race. In J. Glasgow, S. Haslanger, C. Jeffers, & Q. Spencer, *What is Race? Four Philosophical Views* (pp. 4-37). New York: Oxford University Press.

Haslanger, S., & Saul, J. (2006). Philosophical Analysis and Social Kinds. *Proceedings of The Aristotelian Society*, *Supplementary Volumes*, 89-143.

Iannone, A. P. (2001). *Dictionary of World Philosophy.* New York: Routledge.

Iser, W. ([1976] 1986). The Repertoire. In H. Adams, & L. Searle, *Critical Theory Since 1965* (pp. 360-380). Tallahassee: Florida State University Press.

Jeffers, C. (2019a). Cultural Constructionism. In J. Glasgow, S. Haslanger, C. Jeffers, & Q. Spencer, *What Is Race? Four Philosophical Views* (pp. 38-72). New York: Oxford University Press.

Jeffers, C. (2019b). Jeffes' Reply to Glasgow, Haslanger, and Spencer. In J. Glasgow, S. Haslanger, C. Jeffers, & Q. Spencer, *What is Race?* (pp. 176-202). New York: Oxford University Press.

Kant, I. (1998). *Critique of Pure Reason.* (P. Guyer, A. W. Wood, Eds., P. Guyer, & A. W. Wood, Trans.) New York: Cambridge University Press.

Khalidi, M. A. (2015). Three Kinds of Social Kinds. *Philosophy and Phenomenological Research*, *90*(1), 96-112.

Khalidi, M. A. (2023). Natural Kinds. Cambridge: Cambridge University Press.

Kitcher, P. (2003). Race, Ethnicity, Biology, Culture. In P. Kitcher, *In Mendel's Mirror: Philosophical Reflections on Biology* (pp. 230-257). Oxford: Oxford University Press.

Kripke, S. A. (1980). *Naming and Necessity.* Cambridge, MA: Harvard University Press.

Lewis, D. (1978). Truth in Fiction. *American Philosophical Quarterly, 15*(1), 37-46.

Lowe, E. J. (1997). Ontological Categories and Natural, Kinds. *Philosophical Papers*, *26*(1), 29-46.

Luckhurst, R. (2011). Science Fiction. In M. Ryan (Ed.), *The Encyclopedia of Literary and Cultural Theory* (pp. 1257-1266). Malden, MA: Wiley-Blackwell.

Maddy, P. (2015). Class. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (p. 166). New York: Cambridge University Press.

Manley, D. (2015). Fictionalism. In R. Audi, & P. Audi, *The Cambridge Dictionary of Philosophy* (p. 358). New York: Cambridge University Press.

Meiland, J. W. (2015). Category. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (pp. 146-147). New York: Cambridge University Press.

Menzel, C. (2015). Type Theory. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (pp. 1087-1089). New York: Cambridge University Press.

Merriam-Webster. (2024, January 15). *Make-Believe*. Retrieved from Merriam-Webster.com: https://www.merriam-webster.com/dictionary/make-believe

Meynell, L. (2008). Pictures, Pluralism, and Feminist Epistemology: Lessons from "Coming to Understand". *Hypatia*, *23*(4), 1-29.

Meynell, L. (2014). Imagination and Insight: A New Account of the Content of Thought Experiments. *Synthese*, *191*, 4149–4168.

Mills, C. W. (1998). But What Are You Really? The Metaphysics of Race. In C. W. Mills, *Blackness Visible: Essays on Philosophy and Race* (pp. 41-66). Ithaca, NY: Cornell University Press.

Moreland, J. P. (2001). *Universals.* Montreal: McGill-Queen's University Press.

Moritz, R. E. (1914). *Memorabilia Mathematica; or, The Philomath's Quotation-Book.* New York (State): Macmillan.

Oxford English Dictionary. (2024, January 15). Retrieved from OED.com: https://doi.org/10.1093/OED/1043422135

Peirce, C. S. ([1906-8]1998). Excerpts From Letters to Lady Welby. In N. Houser, & C. Kloesel (Ed.), *The Essential Peirce (1983-1913)* (Vol. 2, pp. 477-491). Indianapolis: Indiana University Press.

Plato. (1995). *Phaedrus.* (A. Nehamas, & P. Woodruff, Trans.) Indianapolis: Hackett.

Preus, A. (2007). *Historical Dictionary of Ancient Greek Philosophy.* (J. Woronoff, Ed.) Lanham, MD: Scarecrow Press.

Putnam, H. (1975). The Meaning of "Meaning". *Minnesota Studies in the Philosophy of Science*, 131-193.

Quine, W. V. (1948). On What There Is. *The Review of Metaphysics*, 21-38.

Quine, W. V. (1969a). Natural Kinds. In W. V. Quine, *Ontological Relativity and Other Essays* (pp. 114-138). New York: Columbia University Press.

Quine, W. V. (1969b). Ontological Relativity. In W. V. Quine, *Ontological Relativity and Other Essays* (pp. 26-68). New York: Columbia University Press.

Rankin, S. (2011). Fantasy. In M. Ryan (Ed.), *In The Encyclopedia of Literary and Cultural Theory* (pp. 1054-1058). Malden, MA: Wiley-Blackwell.

Rastier, F. (2004). Type/Token. In O. Houdé (Ed.), *Dictionary of Cognitive Science Neuroscience, Psychology, Artificial Intelligence, Linguistics, and Philosophy* (pp. 434-435). New York: Psychology Press.

Robert the Monk. ([1897]1971). Speech of Urban II. In E. Potts, D. C. Munro, & J. H. Robinson (Eds.), *Translations and Reprints From The Original Sources of European History* (Vol. 1, pp. 5-8 (#44-48)). [Philadelphia] New York: [The Department of History, University of Pennsylvania] AMS Press.

Rosen, G. (1990). Modal Fictionalism. *Mind, 99*(395), 327-354.

Russell, B. (1905). On Denoting. *Mind*, *14*(56), 479-493.

Rust, J. (2021). Max Weber and Social Ontology. *Philosophy of the Social Sciences*, *51*(3), 312-342.

Ryle, G. (2005). Category. In J. Rée, & J. O. Urmson (Eds.), *The Concise Encyclopedia of Western Philosophy* (pp. 72-73). London: Routledge.

Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language* (Vol. 626). London: Cambridge University Press.

Searle, J. (1995). *The Construction of Social Reality.* New York: The Free Press.

Searle, J. (2010). *Making the Social World: The Structure of Human Civilization.* Oxford: Oxford University Press.

Select Committee on the Weights and Measures Act of Great Britain. (1841). *Report of the Commissioners Appointed to Consider the Steps to Be Taken for Restoration of the Standards of Weight and Measure: 1841.* London: W. Clowes and Sons, Stamford Street, for Her Majesty's Stationery Office.

Sellars, W. (1963). Philosophy and the Scientific Image of Man. In W. Sellars, *Empiricism and the Philosophy of Mind* (pp. 1-40). London: Routledge & Kegan Paul.

Spencer, Q. (2019a). How to Be a Biological Raical Realist. In J. Glasgow, S. Haslanger, C. Jeffers, & Q. Spencer, *What is Race?* (pp. 73-110). New York: Oxford University Press.

Spencer, Q. (2019b). Spencer's Reply to Glasgow, Haslanger, and Jeffers. In J. Glasgow, S. Haslanger, C. Jeffers, & Q. Spencer, *What is Race?* (pp. 203-244). New York: Oxford University Press.

Strawson, P. F. (2005). Metaphysics. In J. Rée, & J. O. Urmson (Eds.), *The Concise Encyclopedia of Western Philosophy* (pp. 242-249). London: Routledge.

The King James Bible. (1769 (1989)). The Project Gutenberg. Retrieved from https://www.gutenberg.org/ebooks/10

Thomasson, A. L. (2003). Realism and Human Kinds. *Philosophy and Phenomenological Research*, *67*(3), 580-609.

Toon, A. (2012). *Models as Make-Believe: Imagination, Fiction and Scientific Representation.* New York: Palgrave-Macmillan.

Trogdon, K. (2015). Grounding. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (pp. 430-432). New York: Cambridge University Press.

van Inwagen, P. (2018). *Metaphysics.* Boulder, CO: Routledge.

Villeneuve, D. (Director). (2016). *Arrival* [Motion Picture].

Walton, K. L. (1973). Pictures and Make-Believe. *The Philosophical Review*, *82*(3), 283-319.

Walton, K. L. (1990). *Mimesis as Make-Believe: On the Foundations of the Representational Arts.* Cambridge, MA: Harvard University Press.

Walton, K. L. (2008). Pictures and Hobby Horses: Make-Believe Beyond Childhood. In K. L. Walton, *Marvelous Images : On Values and The Arts* (pp. 63-78). Oxford: Oxford University Press.

Walton, K. L. (2015). Metaphor and Prop Oriented Make-Believe. In K. L. Walton, *In Other Shoes: Music, Metaphor, Empathy, Existence* (pp. 175-195). Oxford: Oxford University Press.

Walton, K. L., & Tanner, M. (1994). Morals in Fiction and Fictional Morality. *Proceedings of the Aristotelian Society, Supplementary Volumes, 68*, 27-50.

Weber, M. ([1921]1978). *Economy and Society: An Outline of Interpretive Sociology* (Vol. 1). Berkeley: University of California Press.

Westerhoff, J. (2005). *Ontological Categories: Their Nature and Significance.* London: (Clarendon Press) Oxford University Press.

Wetzel, L. (2018). *Types and Tokens*. (E. N. Zalta, Editor) Retrieved from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/fall2018/entries/types-tokens/

Williamson, T. (2004). Philosophical 'Intuitions' and Scepticism About Judgement. *Dialectica*, *58*(1), 109-153.

Wolff, A. (Ed.). (2006). *Britannica Concise Encyclopedia.* London: Encyclopædia Britannica.

Wolterstorff, N. P. (2015). Mimesis. In R. Audi, & P. Audi (Eds.), *The Cambridge Dictionary of Philosophy* (p. 671). New York: Cambridge University Press.

Yablo, S. (2021). Say Holmes Exists; Then What? In S. Sedivy (Ed.), *Art, Representation, and Make-Believe: Essays on the Philosophy of Kendall L. Walton* (pp. 345-366). New York: Routledge.

Young, I. M. (2011). *Responsibility for Justice.* New York: Oxford University Press.

Appendix A Key Terms

Representational Objects function as props to equip imaginings.

Props play the specific role demanded of them in a game of make-believe.

Fictionality is the quality of being a fictional truth. **Fictionality** is to imagining how truth is to belief. **Being Fictional** *is analogous to being true*, while *imagining something is analogous to believing something*.

Fictional Truths prescribe, mandate or prompt imaginings in the context of make-believe. Fictional Truths are not truths; they determine fictional contents. *If they match actual truths, then they are also true of the actual world*.

Fictional Worlds are not 'possible worlds' of metaphysics; they are composed of work worlds and game worlds.

Work Worlds are what representational objects provide for generating fictional worlds.

Game Worlds are what participants provide for generating fictional worlds.

Authorized Imaginings are imaginings that generate the game world appropriately and as planned.

Authorized Games are the game worlds generated by authorized imaginings.

The Machinery of Generation describes the components of the system generating fictional truths.

Principles of Generation determine fictional content and guide participants in their imagining.

Direct Principles of Generation are provided by the work world and describe the ways authors stipulate Principles of Generation.

Indirect Principles of Generation are provided by participants and describe indirect ways of guiding imaginings.

Principles of Implication (a.k.a., Principles of Inference) are the subset of indirect Principles of Generation grounded by evidence.

The Reality Principle guides participants to interpret the fictional world like the real world as much as possible.

The Mutual Belief Principle guides participants to interpret the fictional world through the beliefs of the work's author, the community of belief the author belongs to, or the beliefs of the work's intended audience.

Tacit Principles of Generation are the subset of indirect Principles of Generation that are grounded by mere salience.

The Mechanics of Generation describes the operations of the Machinery of Generation.

Endogenous Inference moves from a fictional truth to another fictional truth within a thought experiment.

Exogenous Inference moves from a premise in an argument to a fictional truth within a thought experiment.

Endogenous Problem describes a mistake by an endogenous inference to a thought experiment.

Exogenous Problem describes a mistake by an argument's inference exogenous to a thought experiment.