In our own image: materialism's impoverished image of humanity and what this means for our technological future

Kieran Martin Brayford

Thesis submitted for the degree of Doctor of Philosophy in Philosophy

March 2022

Keele University

TABLE OF CONTENTS:

ABSTRACT

In this thesis, I seek to temper the commonly held faith in our technological endeavours. Recently this has taken the form of seeking to create humanlike technologies. In part one, I argue that progress in this is derailed by the image of humanity advanced by metaphysical materialism and its methodological commitments. I begin by arguing that materialism gains common assent by way of its proximity to the *Promethean attitude* intrinsic to capitalist and communist dogma—the idea that the totality of the natural world can be utilised to improve our quality of life. I then argue that materialism's image of humanity is impoverished as it cannot do justice to several features of human subjectivity—namely, consciousness, free will and selfhood. I then draw upon the difficulties in formulating a reasonable understanding of materialism to further demonstrate that there are no reasonable grounds for endorsing the metaphysic. This culminates in the argument that strong-AI—i.e., that replete with a mind—is impossible in the materialist paradigm. In part two, I explore and evaluate the impact that the Promethean attitude and its weakly-humanlike technological products may have on several fields—healthcare, finance, warfare, work and government—to ascertain the social utility of the attitude. I conclude by warning that such technologies will likely make us wealthier and healthier, but at detrimental cost elsewhere—most prominently, with regard to our agency—and thus, should we not take due care with these technologies, they are capable of creating a future at-odds with the best interests of humankind.

Humanity – Technology – Materialism – Artificial Intelligence – Subjectivity

ACKNOWLEDGEMENTS

My thanks, firstly, go to Prof. James Tartaglia. Not just for his insights, his friendship and his guidance, but for inspiring me to write this in the first place. If it was not for his lectures on the philosophy of mind and metaphysics, the best part of a decade ago, this thesis would have never been written.

Secondly, they go to Dr. Giuseppina D'Oro, who also shares some of the responsibility for inspiring me—her lectures on existential ethics gave me a new understanding of the world that has never left me since.

Thirdly, to my partner, Dr. Amy Johnson. Her resolute exuberance and enthusiasm for my endeavours were (and continue to be) an excellent restorative for when my own reserves ran low.

Fourthly, my family. For teaching me that curiosity is a virtue and that free-thinking is a precious thing, and their ceaseless support in everything that I have done.

And finally, to my good friends, Mark and Mark, who listened to me prattle on.

Kieran Brayford

”Meanwhile, man, precisely as the one so threatened,
exalts himself and postures as the lord of the earth.”

Martin Heidegger
The Question Concerning Technology.
1954. p. 341

***

“There is, indeed, Democritus but we can dispense with him
[…] there is nothing which thinkers of his school cannot
produce out of a bunch of atoms.”

Marcus Tullius Cicero
Tusculan Disputations.
Bk 1, §22

***

“It is not the victory of science that distinguishes
our nineteenth century, but the victory of the
scientific method over science.”

Friedrich Nietzsche
The Will to Power
1888. Aphorism no.466

PART ONE:

WHY MATERIALISM, WHAT IT MISSES OUT AND WHAT IT MEANS TO BE A MATERIALIST

1 IN OUR IMAGE: TECHNOLOGICAL PROGRESS AND THE METAPHYSICS OF HUMANITY

1.1 ALEXA, WRITE AN INTRODUCTION FOR ME

Technology and humankind have always co-existed, but over the last century or so, the world has become increasingly saturated with technology—I will be arguing that this technological progress, under the auspices of metaphysical materialism, is liable to bring us better health, greater wealth, and profound anguish and misery. The widespread belief in materialism, in combination with ever more powerful technologies[1] and huge flows of public and private capital investment[2] has reignited the faith that we will be able to create technologies in our image: a faith that is also stoked by the already existing technologies that gesture towards this goal. Consider the humanoid robots (e.g., Honda's—now retired—ASIMO and Boston Dynamic's Atlas) that we see in highly polished public relations campaigns, or the virtual assistants (e.g., Apple's Siri and Amazon's Alexa) that are found in our pockets and our homes. That these technologies have already seemed to capture something *of the human*, suggests that the incessant technological drive to make technology in our image will come to fruition in the near future.

I argue that this drive is unlikely to be a good thing. My argument is split into two parts. The first part seeks to isolate and evaluate the image of humanity that developers of humanlike technologies use to guide their endeavours. In chapter 1, I present an

[1] For example, consider Moore's Law—"the observation that the number of transistors on integrated circuits have for several decades doubled approximately every two years" (Bostrom, 2014., p. 32). If this progress holds, and techniques to better utilise them (e.g., parallel computing, RISC architectures, etc.) are developed and refined, then computing power may increase considerably in a relatively short amount of time.

[2] In 2015 alone, the UK Invested £11.4 billion in science, engineering and technology (ONS, 2017., n.p.). A year before, in 2014, Google paid $650 million to acquire UK technology firm DeepMind Technologies (Gibbs, 2014., n.p.).

understanding of what this image of humanity is, before then arguing that it is not an image that has arisen from a serious *from first principles* consideration of what it means to be human, but rather it is one that is informed by a materialist metaphysical system that has, in essence, befallen us because of contingent socio-political events and the *Promethean attitude* that such events give rise to. Following this, I explore how this inherited materialism coaxes those loyal to its metaphysical and methodological commitments into endorsing an image of humanity that is thrice impoverished. In chapters 2, 3 and 4, I argue that the methodology of materialism leads it to being unable to grapple with matters of consciousness, free will and selfhood, respectively, which in turn, effectively rules out the possibility of *fully* humanlike technologies in the materialist paradigm. To end the first part, I argue, in chapter 5, that materialism is, in and of itself, not a rational position to hold—not only because of the flaws I explore in chapters 2 to 4, but also because of the difficulties in supplying a reasonable formulation of materialism itself. In the second part, I begin to evaluate the utility of the technologies developed under the flawed materialist paradigm. In chapter 6, I explore the effects of such technologies with regards to healthcare provision and wealth management. In chapter 7, I explore how these technologies will have a detrimental impact on warfare. In chapter 8, I use the exploration of the impact such technologies may have in the realm of work and governance to give credence to that which I call the irony of the Promethean attitude: the idea that the drive to make our lot better by controlling the natural world, ultimately undermines our own agency, thus making our situation worse than it was before. Finally, I end with a warning about the future that our technologies could force us to confront and a suggestion of what can be done to help to limit the chance that this future becomes realised.

1.2 WHAT IT MEANS FOR TECHNOLOGY TO BE CREATED IN OUR IMAGE

We should begin by clarifying what it means to make technology in our image. The drive to make technology in our image manifests itself primarily in the development of artificial intelligences (AIs herein) that incorporate other advanced technologies—such as machine learning and neural algorithms—in order to create *intelligent* technologies. This notion of intelligence here is one that has significant resonances with human intelligence: it is the "possess[ion of] common sense and an effective ability to learn, reason and plan to meet … challenges across … natural and abstract domains" (Bostrom, 2014., p. 4). The inclusion of the notion of common sense here means that we can take the liberty of reformulating it to better draw out its resonances with human intelligence—for a technology to be intelligent it must act as a human would in the context to which it is deployed. All an AI needs to do to be thought of as intelligent is to respond to challenges in a way that is largely analogous to the way in which humans would respond to the same challenge, e.g., by evaluating the challenge presented, drawing up past knowledge and applying it in order to overcome the challenge. This understanding is largely in line with that proffered by the early pioneers into the field—"[the discipline of Artificial Intelligence is] that of making a machine behave in ways that would be called intelligent if a human were so behaving" (McCarthy, 1955., cited in Kaplan, 2016., p. 1)—and our layman understanding of what AIs are. There are two ways in which this humanlike intelligence can manifest itself. Firstly, in a weak sense, whereby technologies *emulate* human capacities, as is the case with "so-called 'expert systems', which are designed to give advice on specialised areas of knowledge" (Crane, 2003, p. 115)—and secondly in a strong sense, whereby human capacities are not emulated, but *possessed* by virtue of "being developed [in line with] a full theory of human

mental processing" (ibid., p. 116).[3] In contrast to weak-AI, this suggests that strong-AI would possess all the subjective faculties that feature in a full theory of human mentality—notably, consciousness, free will and an existentially significant sense of selfhood. Yet, even with this distinction in place, we can note the major similarity between the two—both are designed in line with a humanlike blueprint, both hold the development of humanlike capabilities and faculties as their primary design principle.

1.3 A QUICK SKETCH OF THE METAPHYSICS OF HUMANITY

Naturally, if we are to examine the notion of humanlike technologies, we must first look at the image of humanity that informs their development: we shall do so by offering a quick sketch of the intellectual influences that inform this image. Such an image is rooted in metaphysical materialism, which itself emerges from the shortcomings of the dualism espoused by Socrates, Avicenna and Descartes. The genesis of this dualism is found in Plato's *Phaedo,* specifically in Socrates taking solace in the face of his impending execution with the belief that death is "nothing more than the separation of the soul from the body" (64c). This belief that, in effect, the human is bipartite, the composite of two kinds of *thing*—one that endures after death (i.e., the soul, or the mind) and one that does not (i.e., the body)—which are not only distinct, but are also separable, is picked up later by Avicenna in his flying person argument. The argument runs as follows: imagine that your existence is owed to an omnipotent being, who created you as you are now, but floating a metre above ground with all of your sensory capacities disabled. Despite the fact

[3] This distinction between weak-AI and strong-AI is slightly different in character from the original advanced by John Searle. Searle's version of the distinction holds that weak-AI refers to "the value of the computer in the study of mind" (Searle., 1980., p. 417), i.e., weak-AI allows us to "formulate and test hypotheses [about the mind] in a more rigorous and precise fashion" (ibid), whereas strong-AI refers to the idea that an "appropriately programmed computer really *is* a mind" (ibid). I do not use Searle's distinction in this work for it sees weak-AI primarily as a tool for cognitive science—the distinction I borrow from Crane above better accounts for weak-AI used outside of this context.

that you cannot perceive anything—including your own embodiment—Avicenna suggests that you would still be able to affirm your own existence on the grounds that you would still possess self-consciousness (El-Bizri, 2017., pp. 45 – 46). Such an argument makes a similar distinction to that made by Socrates: "what is affirmed [i.e., the soul, or mind] is distinct to what is not affirmed [i.e., the body, thereby] affirming the existence of the soul as something distinct from the body" (Avicenna, 1027., cited in Goodman, 1992., p. 156). Descartes also agrees with the claim that the human is ultimately composite in nature. The indubitability of the *res cogitans* and the dubitability of the *res extensa,* as discovered via the "general demolition of [his] opinions" (Descartes, 1641., p. 12) during his *Meditations* informs his *Cogito Ergo Sum*—"*I am thinking, therefore I exist*" (Descartes, 1637., p. 127)—which in turn supports the understanding of the human as bipartite, a composite of indubitable mind and dubitable body.

Understanding humanity in such terms is not without issue. As Gassendi was keen to highlight in response to Descartes' dualism, it seems prohibitively difficult to offer an account of causation between two radically different kinds of entity (Gassendi, in Descartes, 1641., pp. 233 – 240). For example, it is difficult to see how the mind, lacking the properties of extended entities, could come to interact *with* extended entities: if the mind is without location, shape or size, its causal relevance for bodily matters is difficult to conceptualise. The materialist's response to this is to move away from dualism and to instead endorse monism—specifically, the form of monism that asserts that all worldly phenomena are owed, in some way, to matter. Moving away from dualism means that a materialist analysis of the mind becomes necessary. Although there are a number of ways that this can be done—for example, by asserting that mental states sit in identity relations with material brain states, as advanced by U. T. Place (1956) and J. J. C. Smart (1959)—

only a specific version of machine functionalism feeds neatly into the project of creating humanlike technologies. Machine functionalism, broadly speaking, argues that mental states are not material states, but are instead defined by the function they perform: particular sensory or psychological inputs are associated with particular behavioural or psychological outputs, and it is this association that informs the functional role that defines a particular mental state (Putnam, 1967., p. 226ff). The value of this for the project of creating humanlike technologies is clear—if mental states are defined by function, then *any* system that can perform the correct function is in possession of mentality: it does not matter *how* or *by what* the function is realised, only that it gets done.[4] This clears the way for the creation of a strong-AI—if the possession of a mind is primarily a functional affair, then all a developer of such a technology needs to do is create a system replete with the appropriate functional efficacies.

Machine functionalism provides much of the conceptual basis of the creation of humanlike technologies, but something more is necessary to guard the project from criticism—as mentioned above, only a *specific* version of machine functionalism is closely associated with the project. Thinking of mentality in fundamentally functional terms seems to neglect the felt, phenomenal quality of mental states—as is suggested by Ned Block's China brain argument (Block, 1978., pp. 275 – 278). Put briefly, one can conceive of a system with human functional equivalence that misses out what it *feels* like to be in a particular mental state: thinking of pain, for instance, in functional terms leaves the *feeling* of pain out of the equation. To head off this criticism, proponents of the technology can augment functionalism with illusionism: "Illusionism makes [the] very strong claim … [that]

[4] This notion is sometimes known as *The Swiss Cheese Principle* (Kirk, 1994., pp. 89 – 92). If performance of the function is all that matters, then it becomes theoretically possible to fashion a mind from anything that can realise the requisite function—even Swiss Cheese.

experiences do not really have qualitative ‘what-it’s-like’ properties” (Frankish, 2016., p. 15). According to illusionism, the *appearance* of phenomenal quality is owed to the fact that we have access to the “covert narrative” (Markkula, 2015., p. 8) of the material goings-on that underpin our mental states. In much the same way as the graphics on our computer screens are the manifest image generated by the thousands of semiconductors found within the machine, our phenomenal experiences are *illusions* that are generated by our ability to access and report on our own material state (Dennett, 1991a., pp. 2019 – 220, 300 – 314 : Dennett, 2017., p. 341ff). This illusionist version of machine functionalism (IMF herein) is of considerable utility to those wishing to develop strong-AI technologies: the project cannot be derailed by a brute appeal to the hard problem of consciousness. All that matters to the project is the development of systems with human functional parity—that the technology may not possess phenomenal consciousness is a non-issue; it is nothing but an illusion anyway. As such, with IMF, we can see the image of humanity that guides the development of humanlike technologies. It is one that sees humankind as a material entity that is distinguished from the rest of the world primarily by what it is that it can *do*. The dualist vision of humankind as a bipartite entity, the combination of the body and the mind is replaced: our functional efficacies are that which make us human to IMF. Insofar as this holds, all that is required to create technologies in our image is the isolation and replication of these functions—create a system with human functional parity, and one has recreated all that is important, in the eyes of IMF, to being human.

1.4 THE HEGEMONIC POSITION OF METAPHYSICAL MATERIALISM

The view of humanity offered by IMF is bolstered by the hegemonic status of metaphysical materialism. In the context of hegemonic materialism, IMF is perhaps the strongest and most plausible image of humanity that one can put forward. It requires nothing that violates materialism—all facets of what it means to be a human can be accounted for in

material terms—and thus it does not offend widespread, common-sense metaphysical sympathies. Yet, this does nothing to explain *why* materialism enjoys its place in the prevailing doxa. Materialism comes *prior* to IMF, insofar as materialism presents the framework on which it is built, so one cannot use the purported sensibleness of IMF as grounds to motivate materialist assent: one ascribes to IMF because they are a materialist, not vice versa. Neither can one attribute materialism's hegemonic position to academic philosophy. Although it features heavily in contemporary philosophical discourses, this influence has largely been confined to the academy—such discussions have not been altogether influential on the common-sense positions of ordinary people, or those working to develop humanlike technologies.[5] As such, if we are to understand *why* materialism has become embedded in the prevailing doxa, we must look to some factor that sits beyond the academy which acts to imbue materialism with popular appeal—something that accounts for its widespread assent.

A clue about what this might be can be found if we look at the history of materialism. If we ask ourselves *where and when* materialism has enjoyed its highest levels of assent, we find that there are two answers to this question. The first is in the Anglo-Atlantic countries, where analytic philosophy is dominant—in these countries, materialism currently enjoys high levels of assent, as evidenced by its prominence in philosophical literature. The second is in the Eastern Bloc during the 20th century, where materialism was an integral part of the prevailing *doxa* of these societies also. These instances of high assent are no accident. Rather they arise from the conditions of these societies: materialism draws its dominance from its interplay with the politico-economic situations of these societies, and

[5] This may not be the case for those who are religious—it seems sensible to assume that the religious would not hold materialism in such high esteem.

as such, we can distinguish between two kinds of materialism—capitalist materialism and communist materialism—which, though broadly alike in content, differ with regards to how they come to enjoy their hegemonic status. We shall begin by exploring how capitalist materialism acquires its widespread assent before then giving similar treatment to communist materialism.

1.5 CAPITALIST MATERIALISM I: A BRIEF SKETCH OF THE CAPITALIST SYSTEM

To understand how capitalist materialism attains its assent, we must offer a brief sketch of the capitalist system and its history. The system itself can be explained through the interplay between three broad phenomena: technology, utilisation of natural resources and economic growth. The foundation of the system is found in the utilisation of nature by self-interested actors using technological apparatuses in order to reach a position of self-sustenance through engaging in "the propensity to truck, barter and exchange one thing for another" (Smith, 1776., p. 17). Consider the example from Adam Smith, of the arrow-maker in a tribe of hunters—by taking materials afforded to them by nature and working them with tools, in accordance to their "readiness and dexterity" (ibid., p. 19) the arrow-maker can create a surplus of arrows that can be exchanged for the surplus created by other specialised craftsmen: the arrow-maker may exchange his surplus for the surplus of the hunter, swapping arrows for meat. This *technique*, of dividing labour amongst a populace can be extended to create greater surpluses to exchange for a larger share of the surpluses created by other people. If a team of arrow-makers—one focussing on the fletching, the other the shaft and the last the head—combine their efforts, they are able to create more arrows than a lone arrow-maker could, not just because of the time saved "passing from one sort of work to another" (ibid., p. 12) but also because of the increased skilfulness of the worker who can focus their efforts on one single task. These gains can be further increased by exchanging the created surplus for technological apparatuses that improve the

efficiency of the production process: labour can be “facilitated and abridged by the application of proper machinery” (ibid., p. 13). As such, the division of labour can take an even finer-grained form, whereby one worker is responsible for only a very small part of a finished product, thus increasing the surplus created further.

This basic pattern of economic growth—i.e., the application of technology to both allow for the utilisation of natural resources, and to make such more efficient—has been the driving force behind the several industrial revolutions that precipitated the contemporary capitalist system. Each time a new technology that allows for more efficient harnessing of natural resources is developed and deployed to society, an increase in economic growth is triggered. The first of these revolutions occurred between the mid-18th and the mid-19th centuries and began with the invention of a number of productive technologies “that “heralded the advent of a mechanized production system” (Chang, 2014., p. 53), and ended when “the invention of the steam engine … ushered in mechanical production” (Schwab, 2016., p. 7). The second industrial revolution occurred some decades later, spanning the end of the 19th century and the early 20th, and was driven by “new technologies … developed through the systematic application of scientific and engineering principles” (op cit., 2014., p.65), thus allowing for the rapid replication and deployment of these technologies. These technologies, “fostered by the advent of electricity” (op cit., 2016., p. 7), culminated in the invention of the assembly line, which allows for the efficient and fine-grained division of labour that still endures in today’s manufacturing processes. Following this, the development and deployment of information and communications technologies, and efficient transportation technologies during the latter half of the 20th century spurred a further revolution in capitalist productive capacities, whereby the division of labour was able to expand so that the production process was internationally

integrated, thus enabling today's globalised, interconnected economy (op cit., 2014., pp. 96 – 98). As we progress through the 21st century, a number of emergent technologies—e.g., (weak-)AI, distributed cyber manufacturing, machine-to-machine communications, etc.—are likely to be sufficiently disruptive as to constitute a fourth industrial revolution (op cit. 2016., pp. 7 – 8), again unlocking a more efficient and lucrative utilisation of our natural resources.

1.6 CAPITALIST MATERIALISM II: ECONOMIC GROWTH AND QUALITY OF LIFE

Each of these revolutions have accelerated economic growth—this has manifested itself in an increase in the purchasing power of the average person in developed societies. This rise in purchasing power, in turn, has led to a diversification of "the structure of consumption" (Piketty, 2014., p. 87). Whereas in early industrial society, much of the average worker's income was spent on subsistence, consumption patterns have changed to "a much more diversified basket of goods, rich in manufactured products and services" (ibid). This diversification of consumption habits is indicative of "the spectacular increase in standards of living since the Industrial Revolution" (ibid). Now that workers are free to spend a smaller portion of their overall income on subsistence, the remainder can be used to achieve other ends, whether this be through direct spending, or through the levying of tax that is used to fund public works and services. Technologically driven economic growth has not just improved the purchasing power of the populaces in developed societies; it has also caused gains in quality of life elsewhere, perhaps most importantly in the realm of education. The increasing efficiencies of the manufacturing and agricultural sectors has led to a considerable swing away from workers being employed in these sectors and towards employment in the service sector (ibid., pp. 90 – 91). The service sector represents a wide range of roles—from retail and hospitality through health and education to government administration—and thus requires the existence of an educated and skilled populace from

which to draw its workers. This need for a skilled populace has precipitated the expansion and democratisation of education seen in developed economies following the Second World War.

The expansion of education was not just necessitated by technological advancement, it was also *enabled* by technological advancement. Consider the technologies that allow for the recording and mass distribution of the knowledge base of humanity, i.e., "writing, printing and electronic media" (Pinker, 2018., p. 233), and the role that these play in education. Not only do they allow for the development of a standardised curriculum that can cater to the needs of society, but it also allows for a high degree of autodidacticism. The benefits of this expansion of education, according to Steven Pinker, are twofold. Not only is there a direct benefit from an educated populace insofar as higher levels of education are associated with the uptake of liberal cosmopolitan values—"educated people … are less racist, sexist, xenophobic, homophobic and authoritarian" (ibid., p. 235)—but there is also a secondary benefit as an educated populace is able to aid the development of the technological base of society. With a highly educated populace comes the potential to develop advanced technologies, e.g., (weak-)AI, genetic engineering, cyber manufacturing systems, etc., that can be put toward humanitarian ends and improvements in quality of life (ibid., pp. 330 – 333). For example, (weak-)AI could help to develop novel treatments to disease, genetic modification of crops could improve yields, etc. (ibid., p. 331ff). As such, technology is envisaged as something more than machinery or technique—it is instead that which allows us to escape our condition of brutishness (Brayford, 2020a., p. 527)—and thus, is viewed through an optimistic lens.

From this we can see that technological advancements have, overall, had a positive impact on the quality of life of those in advanced economies—the question of how this acts to incorporate materialism into doxa remains. Let us recall that an appropriate gloss on the materialist's position is that everything is either material, in and of itself, or otherwise supervenient upon the material. This means that the natural world, and all of it goings-on are also material, or supervenient on the material. This is notable, because the material world is one that can be manipulated by technologies—should we look at any technological advance we note that their functions are realised through the manipulation of the material world (e.g., the intricate manipulations of electrons that underpin digital technologies). This means, because materialism is a monistic metaphysic, that there is no theoretical limit to the manipulation of our world: if everything is matter, and we can manipulate matter with our technologies, then it follows that there are no theoretical limits on our ability to use our world to serve our ends. Therefore, the technological utilisation of nature has no metaphysical limits—everything is able to be bought under our dominion, provided we have the technology available to do so. Yet, it is not just the utilising of nature that is free from metaphysical limits under materialism—as many of the gains in our quality of life are intertwined, in one way or another, to the technologically driven utilisation of our natural resources, then these gains are also without metaphysical limit. Arguably, it is from here that capitalist materialism draws its assent—not from the philosophical robustness of its premises, or from any argumentative panache, but from materialism's implied promise of a better future.

1.7 COMMUNIST MATERIALISM I: HEGEL, MARX AND HISTORY

The dominance of metaphysical materialism in the communist countries in the 20th century is owed to a different set of circumstances than that of capitalist materialism, though there is indeed some overlap between the two. To understand from where communist

materialism draws its assent though, the work of Karl Marx and his patron, Friedrich Engels must be considered. The ideas of Marx and Engels formed much of the philosophical underpinnings of the ideology of the Soviet Union and its satellite states and thus played a significant part in installing materialism in the doxa of communist societies.

Karl Marx's materialism was rooted in his admiration of the Atomists of ancient Greece—his doctoral thesis explored the differences between the Democritean and Epicurean philosophies of nature—but it was Marx's critique of the Hegelian theory of history that popularised his version of materialism. Hegel's understanding of history draws upon his own notion of *Geist*, or Spirit. Geist, for Hegel, is an immaterial being: whereas material beings are "determined by a force outside of [them]" (Hegel, 1840., p. 20), Geist is "that which has center in itself … a Being-by-itself" (ibid)—it relies on nothing other than the consciousness of itself for its existence. It is "self-sufficient" (ibid) as it needs nothing else other than itself for its instantiation. From this, we can see that Geist is a kind of consciousness, specifically, a "consciousness that *has Reason"* (Hegel, 1807., p. 265). What Hegel means by this is that Geist shows consistency in its activities; like nature, its actions are regulated in accordance with some manner of laws or maxims. But, unlike nature, which has to follow natural laws that it itself has not created, Geist—by virtue of its self-sufficiency—is able to impose the laws that regulate its activities *upon itself*. Geist is self-regulating: the constraints to its actions are self-imposed. The precise nature of these constraints is reflective of Geist gradually gaining consciousness of itself and its self-sufficient nature: "it is the judging of its own nature, and at the same time it is the activity of coming to itself, of producing itself, making itself actually what it is in itself potentially" (ibid., pp. 20 – 21). As Geist comes to grapple with its potentialities, as thought undergoes "the efforts … to comprehend itself" (de Boer, 2009., p. 52), the rules that it imposes upon

itself shift; the rules that Geist follows when it has a slender grasp on its own nature are different to those in place when its grasp of itself becomes firm.

It is Geist, in Hegel's view, that determines the unfolding of history: "world history … is the exhibition of the Spirit, the working out of the explicit knowledge what it is potentially" (op cit., 1840., p. 21). The precise configuration and character of a particular historical epoch is a manifestation of the Geist coming to know itself, or, in other words, the configuration of a society is owed to the society's own collective understanding of its own nature. A society's self-understanding—its grappling with its Geist—is revealed, in Hegel's view, not only in "its religion, its art, and philosophy, … [and] its thoughts and imaginings … in general" (ibid., p. 48), but also in the way that its self-understanding informs the organisation of that society. The character of a society shifts at the point at which "the particular determination of freedom on which [it] relied is at odds with the principle of freedom as such" (ibid., p. 53), i.e., "when a civilization has exhausted the possibilities opened by [their] particular determination of freedom" (ibid). When a society's understanding of itself "no longer enhances [its] well-being" (ibid), we see the unity of the society begin to fragment. A period of civil disunity occurs that necessitates a "rebirth and renewal" (Cohen, 2000., p.19): a recalibration of society's governing ideals is needed to transcend the fragmentation of society and restore its unity. This renewal is "often associated with the insight and struggle of a great man … an Alexander, a Caesar, a Luther, or Napoleon" (ibid) who overcomes the fragmentation of society by shifting the paradigm of ideals that informs society, which then triggers a societal recalibration: "in a world of roles in disarray, [he] fashions a new role for himself and a new script for others" (ibid).

By acting in a manner that transcends society's self-understanding, these individuals are able to put the Geist in touch with a deeper understanding of itself, thus recalibrating society in a manner that allows the possibility of greater self-regulation, effectively extending freedom to greater sections of society. For Hegel this is evidenced retrospectively, by examining the unfolding of history. He suggests that in early civilisations—"in the world of the ancient Orient" (op cit., 1840., p. 48)—where society's understanding of its ability to self-regulate is slight (or, in Hegelian terms, where the Geist is not yet fully conscious of its freedom), we find a situation where "people do not yet know that the Spirit … is free … they only know that *one* person is free" (ibid), but as the understanding of self-regulation deepens, as individuals emerge that shift society's understanding of itself, we see the extension of freedom—self-regulation—to greater portions of society, culminating, in Hegel's view, with the "Germanic peoples, [who] through Christianity, … came to the awareness that *every* human is free by virtue of being human" (ibid). From this, we can see that Geist's role in the development of history is that of setting the "mode of thought that underlies the efforts of successive civilizations both to organize themselves in a rational way and to comprehend the principle of self-organization" (op cit., 2009., pp. 52 – 53).

Marx disagrees with this understanding of historical development. For Marx, Hegel's idealism—which holds consciousness to be "the demiurgos of the real world" (Marx, 1873. p. 11)—mistakes the relationship between the idea and the world. Whereas Hegel's theory of history holds that the qualities and characteristics of a society are an expression of the ideas of that society, which take "external, phenomenal form" (ibid), Marx holds that the inverse is true. Society, to Marx, is not configured in line with the adherence to self-imposed maxims. Rather, because "the ideal is nothing else than the material world

reflected by the human mind" (ibid) the ideas of a society—a society's Geist—are informed by the *material conditions* of the society. In Marx's words: "it is not the consciousness of men that determines their existence, but their social existence that determines their consciousness" (Marx, 1859., p. 220). The world is not an expression of an idea, but rather our ideas are expressions of the world. History therefore, is "controlled not by conceptions of man but by material ends and means" (op cit., 2000., p. 22).

The configuration of a society owes no debt to society's self-understanding for Marx because societal configuration corresponds to the productive capacities of that society: "in acquiring new productive forces men change their mode of production; and in changing their mode of production … the way of earning their living, they change all their social relations" (Marx, 1847., p. 95). Contra Hegel, Marx holds that qualitative changes between historical epochs is owed to the development and deployment of technologies to society, because these technologies bring with them a shift in the mode in which humankind labours—"the hand-mill gives you society with the feudal lord' the steam-mill, society with the industrial capitalist" (ibid)—and it is the mode of labour that governs one's position in society: the proletariat are those that perform labour to subsist, and the bourgeoise are those that need not to do so. Moreover, productive technologies do not just inform the configuration of society but they also determine "principles, ideas and categories" (ibid) found in that society: the cultural products of a society, in Marx's view, do not represent the progress of the spirit's knowing of itself, rather they represent the current moment of technological progress. The progress of history, according to Marx, therefore is the progress of technology insomuch that shifts between epochs are precipitated by shifts in productive capacities.

This said, for Marx, there is nothing inevitable about the flow of one epoch into the next; history does not unfold in line with some manner of metanarrative. Rather the shift between epochs is *contingent*—shifts do not *only* occur through the development of new technologies (though the deployment of technologies does cause a necessary shift in societal configuration). They can also be bought about in response to conditions that productive technologies bring with them—human endeavour to escape their situations can also be sufficient for epoch change. The deployment of mechanised productive technologies, in Marx's view, not only is a "competitor to the workman himself" (Marx, 1867., p. 264) in the sense that their "means of livelihood [has] been destroyed by machinery" (ibid) that can perform their labour in their place thus rendering them without income, but is also the cause of much working-class suffering: "the physical and mental degradation, the premature death [and] the torture of over-work" (ibid., p. 166) are in Marx's view, a consequence of the system of production ushered in by these productive technologies.

As such a system of production matures, so to "grows the mass of misery, oppression, slavery, degradation [and] exploitation" (ibid., p. 379), but alongside this, according to Marx, so "too grows the revolt of the working class" (ibid). Marx shares the same broad schema of historical change as Hegel—society must fragment due to the insufficiencies of its organisation before it can then be reorganised and reunified. It is precisely the growth of revolt in the working-classes that represents the stage of fragmentation. But here Marx again breaks with Hegel—the society cannot change its organisation via the arrival of a great man, rather, in Marx's view the recalibration of society must emerge from the *"*free conscious activity" (Marx, 1844., p. 328) of the working classes, i.e., through *praxis*. In Marx's view, the fragmented society can be reunified "*only* in a *practical* way, only

through the practical energy of man" (ibid., p. 354). The working-classes must re-focus their capacities for work that have been co-opted by capitalism as alienated labour away from such and towards bringing about communism. It is not a great man that will reunify society in Marx's view, for the actions of a single individual can do little to recalibrate society in and of themselves, because a shift in epoch caused by humanity (as opposed to technology) must be effected by collective action: humans must work to change the material basis of their society—it is only the collective actions of mankind that can bring forth the form of society that transcends its fragmentation.

1.8 COMMUNIST MATERIALISM II: ENGELS AND DIALECTICAL MATERIALISM

Karl Marx and Fredrich Engels are often quoted in tandem, as if to suggest that there is total unity in their thinking, but this is not the case. In reality, Marx's thinking was subverted by Engels—"the first revisionist, the point of origin of the bifurcation of the world Marxist movement" (Levine, 1975, p. xv). Inspired by Marx's theory of historical change—historical materialism—Engels wished to 'complete' Marx's work by "mov[ing] beyond the realm of human history" (ibid., p. 142) and "demonstrat[ing] that the operational laws of nature itself" (ibid) worked in accordance with the dialectic present in Marx's historical materialism. To do this, Engels revisits the "mere laws of *thought*" (Engels, 1883., p. 26) developed by Hegel in his *Science of Logic*—"[1] the law of the transformation of quantity into quality …; [2] of the interpenetration of opposites; [and] .. [3] of the negation of the negation" (ibid)—and demythologise them by applying them to the natural world, thereby recasting them as the dialectical laws of his own *dialectical* materialism.

Each of these laws, in Engels' view, can be substantiated by the material world as revealed by the natural sciences. According to Engels' the first law we can see at work at an atomic

level. He suggests that the splitting up of a body "into smaller and smaller portions" (ibid., p. 28) is met with no substantial change of quality in the objects, until one reaches the molecular level, at which division is necessarily accompanied by "a complete change of quality" (ibid). The atoms that constitute the molecule are qualitatively distinct from the molecule itself, and as all is made fundamentally from atoms, changes in quality of a large body, in Engels' view, must correspond to quantitative changes on an atomic level. The second law, Engel's sees instantiated in the mechanistic movements of atoms—which due to the "law of the indestructibility and uncreatibility of motion" (ibid., p. 38) are necessarily characterised by "the old polar opposites of *attraction* and *repulsion*" (ibid). The movement of a body can be understood either as attraction *towards* another body, or the *repulsion* away from another. The final law—the negation of the negation—Engels sees instantiated in the natural world in instances where an entity transcends its current condition, most obviously in the development of living creatures. For example, Engels speaks of "Butterflies … spring[ing] from the egg by the negation of the egg" (Engels, 1878., p. 188) only to undergo another negation after they mate, lay their eggs and die (ibid).

From this we can see the distinction between the thought of Marx and the thought of Engels. For Marx "the field of materialist enquiry was history and society" (Eagleton, 2016., p. 62), but in his application of the laws of the dialectic to the natural world, we see that Engels "was a materialist in a standard philosophical sense of the term" (ibid., p. 61). Engels was a "totalist" (op cit., 1975., p. 142)—his understanding of the natural world "can be reduced to the claim that everything that was, was matter and its motion" (ibid., p.

145).[6] Moreover, Engels' dialectical materialism, because of its insistence that all phenomena are reducible to atomic matter and its movements that necessarily operate in accordance with dialectical laws, is necessarily deterministic: all is "solely dependent upon the blind mechanics of nature" (ibid). Engels' efforts to apply dialectics to nature excised praxis from Marx's theory of history; gone was "the Marxian vision … [of] man who acted" (ibid., p. 152) and in its place we find the "cold, unremitting and remorseless system" (ibid., p. 145) of materialist monism. For Engels, contra Marx, human efforts to change society, to engage in praxis to bring about communism, had little impact on the unfolding of history for "causation originated and flowed from the physical … the same laws that governed the physical universe also governed the social universe" (ibid). The shift away from capitalism into communism becomes a metaphysical necessity under Engels' patronage, guaranteed by the determinism of a universe following the laws of the dialectic.

1.9 COMMUNIST MATERIALISM III: LENIN AND STALIN

Engels' deterministic reading of Marx was deeply influential on Vladimir Lenin and Joseph Stalin. Following Engels, Lenin also endorsed the primacy of the material world, and thus also understood all phenomena and all historical events to be a consequence of the world's unfolding in line with the laws of the dialectic—"the world is matter moving in conformity to law" (Lenin, 1908., p. 169). The shift to communism was inevitable in Lenin's view, as it was the "necessary result of the development of the productive forces in modern society" (Lenin, 1896., p. 19)—and this belief seems to be the justification Lenin held when pursuing violent revolution in order to bring about communism. The "concrete analysis of each specific historical situation" (Lenin, 1916., p. 316) led Lenin to the belief

[6] Here we can see the influence of the Ancient Greek atomists on Engels—the assertion that all there really is, is atoms and their movements is shared by Engels, Leucippus and Democritus alike.

that to bring about communism, without first undergoing capitalist development and enduring the hardships it brings to the working-classes, a violent revolution led by a revolutionary vanguard was not only necessary but was the fulfilment of a course of action that was always already preordained by the deterministic operations of the material world. The installation of Lenin's "dictatorship of the proletariat" (Lenin, 1917., p. 402) and the subsequent "suppress[ion] of the bourgeoisie" (ibid) was in his view, simply the hastening of the prophecy written by the dictates of Engels' dialectical laws. It was he, and his revolution, that was metaphysically mandated to bring about revolution in order to then reunify society by reorganising it in a communist configuration.

The precise manner in which Engels' dialectical laws supposedly acted to necessitate Lenin's revolution was set out in detail in Stalin's work on dialectical materialism. For Stalin, the laws of the dialectic reveal Lenin's revolution as a natural and inevitable development, insofar as it is mandated by and reflective of these laws. Stalin notes that in the world as understood by the natural sciences, "imperceptible and gradual" (Stalin, 1938., p. 8) quantitative change begets sudden qualitative changes, changes that occur "rapidly and abruptly, taking the form of a leap from one state to another" (ibid). For Stalin, this law mirrored the social situation surrounding the revolution: as the peasantry faced "disintegration" (ibid., p. 13) and the proletariat underwent "development" (ibid)—i.e., as the numbers of the peasantry dwindle and shifted to the proletariat—the revolution came to be necessitated as the embodiment of the rapid and abrupt changes foretold by the laws of the dialectic. In Stalin's words: "if the passing of slow quantitative changes into rapid and abrupt qualitative changes is a law of development, then it is clear that revolutions made by oppressed classes are a quite natural and inevitable phenomenon" (ibid., p. 14). Similarly, Stalin also sees the revolution as reflective of the law of the

interpenetration of opposites. Like Engels, he suggests that "dialectics holds that internal contradictions are inherent in all things and phenomena of nature" (ibid., p. 11) and that development naturally takes its form in a struggle "between the old and new, between that which is dying away and that which is being born" (ibid., p. 11)—a point that he underscores by approvingly citing Lenin: "Development is the 'struggle' of opposites" (Lenin, 1908., cited in Stalin, 1938., p. 11). Through this lens, Stalin sees the revolution not only as necessitated by the current conditions brought about by the dialectical laws, but as necessary for the transition into communism. Peaceful reform is not an adequate reflection of the struggle he sees as inherent in all things—only violent revolution and the suppression of the bourgeoisie can fulfil the mandate given by the dialectical laws: "one must pursue an uncompromising class policy, not a reformist policy of harmony of the interests of the proletariat and the bourgeoisie" (ibid., p. 14).

We can see from this how materialism came to prominence in the countries under the political influence of the USSR: materialism formed part of the founding myth of the communist societies. Communism owed its existence to the supposed truth of dialectical materialism, for it was dialectical materialism that justified the revolution as inevitable. The transition to communism, through the lens of dialectical materialism, was always on its way—a natural product of the operation of the dialectical laws. The common assent to materialism in these societies was achieved in two ways. For devout communists, materialism contained within it the promise of a better life, free from the kind of exploitation that they believed to be prevalent in capitalist societies. For those sceptical of communism, the state assured their tacit assent to materialism through the "power of censorship and threat of punishment" (Barber, 1979., p. 142).

Materialism, because of its role as a founding myth of the state, was key to maintaining the power of the various communist parties in their respective countries. Its inverse, idealism, was held to be a "fashionable *reactionary* philosophy" (op cit., 1908., p. 130) by Lenin who, moreover, saw dialectical materialism as "precisely the rejection of the methods of idealism and subjectivism" (Lenin, 1894., p. 182). This hostility toward idealism was mirrored by Stalin, who viewed "Marxist philosophical materialism [as] … the direct opposition of philosophical idealism" (op cit., 1938., p. 15). In this context, there was no toleration of anti-materialist views—they were denounced for their supposed anti-revolutionary sentiment. This led to a situation where "intellectual neutrality and academic autonomy soon ceased to be options [and] tolerance of non-Marxist intellectuals … was replaced by the demand for unequivocal commitment to the official world-view" (op cit., 1979., p. 141). In the context of this restriction of academic liberty "what Soviet philosophers produced looked far more like quotatology … than philosophy" (Bocheński, 1967., p. 2)—philosophy was reduced to producing only that which could be assimilated with dialectical materialist doctrine; any public deviation from dialectical materialism was simply not tolerated, and so, it found itself firmly placed within the prevailing doxa of communist societies.

1.10 ENVIRONMENTAL DEGRADATION AND THE PROMETHEAN ATTITUDE

If we strip communist materialism of its revolutionary zeal, we find that it shares an important similarity with its capitalist counterpart—both hold technology, and its ability to manipulate the natural world to be the harbinger of greater human flourishing. Both see technology as the driving force behind a more equitable, more civilised and generally happier human existence: the major difference between the two lies in their preferred administrators of technology. For capitalism, technology developed and deployed in line with the logic of the free market naturally leads to gains in quality of living—communism

holds that these beneficial effects of technological development are best realised if the dictatorship of the proletariat, not the free market, oversees technology's development and deployment. There is though, a tension here between the manipulation of the natural world for humanitarian ends, and the *effects* of such manipulation.

As Hans Jonas argues: "modern technology, informed by an ever-deeper penetration of nature and propelled by the forces of market and politics, has enhanced human power beyond anything known or even dreamed of before" (Jonas, 1979., p. ix). Humankind, with each major advance in its technological capabilities is able to bring more and more of the natural world under its dominion, "bending circumstances to [our] will and needs" (ibid., pp. 2 – 3) and putting nature to work in increasing our quality of life. But with this gradual process of bringing the natural world under human dominion comes a set of "slow, long-term [and] cumulative" (ibid) issues, broadly described by Jonas as the "raping of nature" (ibid., p. 2). By extending its dominion over the natural world, humanity has changed, in Jonas' view, "no less than the whole biosphere of the planet" (ibid., p. 7). These changes have profound consequences for the natural world: a reliance on fossil fuels to power our technologies has caused a rise in atmospheric CO_2 levels which "increase[s] the risk of irreversible climate change, such as the loss of major ice sheets, accelerated sea-level rise and abrupt shifts in forest and agricultural systems" (Rockström, et al., 2009., p. 473). The repurposing of wildlands into agricultural lands not only contributes to these emissions, but also has its own impact: the rate of biodiversity loss has "accelerated massively" (ibid), and the use of fertilisers causes a degradation of aquatic ecosystems as they leach into waterways, thus making "anoxic ocean events more likely" (ibid., p. 474). As humanity tightens its grip over nature, these ill effects increase in magnitude—by seeking a better life through the harnessing of nature via technology, mankind has caused a situation where

its own survival comes under threat. If the trend of ecological decline continues, humankind risks making its own environment uninhabitable, "for the future of all nature on this planet [is] a necessary condition of man's own" (op cit., 1979., p. 136).

The problem is clear—the technology-driven utilisation of our natural world, bolstered by our politics, a wish to improve our standard of living, and a metaphysical system that poses no conceptual barrier to the extent that we can harness nature, reaches its zenith in jeopardising not only the sustainability of our standard of living but our continued survival as a species. Yet, the same kind of thinking that has engendered this situation also sees the problem as surmountable. Recall the role materialism plays in this system—by presenting everything as material, it holds that there are no limits to our capacity to utilise nature for our ends: everything is matter, matter can be manipulated by technology, therefore we can manipulate matter with our technologies to improve our standards of living. This same logic can also be applied to the problem of ecological degradation: if everything is matter, then ecological degradation is a *material* problem and thus can be overcome via technology—an attitude that is also encouraged by a pervasive optimism towards technology in response to its role in improving the quality of our lives. Metaphysical materialism, combined with an optimistic attitude towards technology, gives rise to the *Promethean attitude*, i.e., the "unlimited confidence in the ability of humans and their technologies to overcome any problems" (Dryzek, 2013., p. 52).

This Promethean attitude has long been present in developed societies, collecting more adherents as attempts to improve our lot bore fruit—capitalism and communism, the two dominant politico-economic systems of the previous centuries, shared a commitment to the Promethean attitude, as evidenced by their faith in metaphysical materialism and the power

of technology to increase our wellbeing. Indeed, I think that the social utility of the Promethean attitude (and by extension, its constituent parts—materialism and technological optimism) is the major factor informing its hegemonic status. The Promethean attitude became part of the prevailing doxa because it *worked*—it has been a useful tool in escaping the brutishness of our existence. The gains in the quality of our lives made over the past epoch are owed, in no small part, to the Promethean attitude's empowerment of humanity and its technological products. But, as we have seen above, the utility of such an attitude may be beginning to wane—the faith that our world is endlessly manipulatable, and that our technologies assure the resolution of our problems has led to the degradation of our environment and a threat to our survival—if the other products of the Promethean attitude are similarly problematic, then it may signal the need for a shift in our thinking.

To gauge whether a shift in thinking is necessary, we must present an evaluation of the Promethean attitude and the related endeavour of creating technologies in our image. This will begin by an evaluation of the materialism that is associated with the attitude. The remaining chapters of part one will focus both on how materialism and its methodological commitments lead the project of creating humanlike technologies to grapple with an impoverished image of humankind, and how formulating materialism itself is fraught with difficulty. We will return directly to the Promethean attitude in part two. There, we will evaluate the practical effects of the attitude on a number of different fields—healthcare, finance, warfare, work and government—so we can understand how the Promethean drive to create a better existence is liable to do the opposite.

2 THE IMPOVERISHED IMAGE I: CARE, CONSCIOUSNESS AND THE VIEW FROM NOWHERE

The image of humanity advanced by IMF (illusionist machine functionalism)—i.e., that where the human is thought of in primarily functional terms—serves as the blueprint for the creation of humanlike technologies. Though such an image is perhaps the most natural and consistent view of humanity for a materialist, I will argue that it is an *impoverished* image of humanity. In this chapter I will demonstrate how the methodological commitments of IMF—employed, ultimately, out of deference to materialism—are insufficient to understand human consciousness. I shall give similar treatment to the notions of free will and selfhood in the following chapters. Here, I shall first isolate the goals and the methodology of IMF, before noting how these rely on the impossible removal of anthropocentrism to investigative endeavours in order to achieve detached neutrality. I shall look at how this neutrality is supposedly achieved before drawing on the work of Wilfred Sellars, W. V. O. Quine, Richard Rorty and Martin Heidegger to demonstrate how such neutrality is impossible. After showing arguments against such a view from neo-realists Graham Harman and Maurizio Ferraris to be flawed, I then employ Heideggerian hermeneutics and his notion of *aletheia* to demonstrate how investigative methodology can distort the object of investigations, before finally arguing the image of humanity advanced by IMF misunderstands consciousness precisely because it relies on a methodology that is unfit for the job of understanding consciousness.

2.1 THE METHODOLOGY OF IMF: SMOKE, MIRRORS AND MAGIC

If we are to offer a critique of IMF's view of humanity, then it is wise to first examine its motivations and methodology. IMF is frank with regards to its aims. Imagine one is sitting in a theatre, watching a magician perform their tricks—as they saw their assistant in half, it

seems as though they have really been split in two, but this is nothing but an illusion. If we leave our seats and investigate the magician's equipment, we would find out how they achieved the illusion: the lower box contains a set of false legs and the assistant contorts themselves into the upper box. The role of IMF is broadly similar to that of the inquisitive audience member. In much the same way that the inquisitive audience member finds out how the magician's trick is performed, the cognitive scientist must find out about how the illusion of consciousness can be explained: "[it is] our *burden* to explain *how the 'magic' is done*" (Dennett, 2017., p. 65ff). More specifically, IMF seeks to "provide the details of how the brain manages to create the illusion of phenomenality" (ibid., p. 66)—it wishes to explain the illusion of consciousness in terms of the functional mechanics from which the illusion arises. Thus, the project of IMF is built upon the assumption that there is a dichotomy between two different images—a *manifest* image that leads us to the false belief that consciousness is real and a *scientific* image that explains the manifest image in the terms of that which is empirically available to us. The goal of IMF is to articulate this scientific image.

2.2 ASK AND IT WILL BE GIVEN TO YOU: THE PHENOMENOLOGY OF ANOTHER

The goal of IMF is to uncover the hidden science that produces the manifest image of phenomenality. Our next move is to examine the methodology by means of which it seeks to achieve this goal: a goal is folly unless we have some idea as to how we are to achieve it. The methodology employed by IMF, according to Dennett, is *heterophenomenology*. As its name suggests, heterophenomenology is a variety of phenomenology, yet with enough differentiating it from the classic conception of phenomenology to warrant its status as a stand-alone methodology. Whereas classical phenomenology takes direct experience as the object of its investigation, heterophenomenology takes the experiences of *others* as the

object of its investigations: heterophenomenology is the "phenomenology of another, not oneself" (Dennett, 2003., p. 19). Yet, this description tells nothing of what this entails, practically speaking—how does one *do* heterophenomenology?

By Dennett's own admission, it is simple: "most of the method is so obvious and uncontroversial that some scientists are baffled that I would even call it a method" (ibid., p. 20). Heterophenomenology is a collaboration between the experimenter and their subject—the experimenter asks the subject about what they believe to be true with regards to their conscious experiences and then *listens* to these reports (ibid), building up a bank of the participant's beliefs about their subjective experiences and thus isolating a phenomenon—the subject's (illusionary) consciousness—that is poised to be explained by the work of cognitive scientists. Dennett's assessment of this methodology seems fair—there does not seem to be much controversial here. Provided that the participants of the experiment have a sufficient grasp on language as to avoid any miscommunication and the experimenter has no reason to doubt the truthfulness of the reports, then there seems to be no reason that we cannot take these reports as accurately representing the participant's manifest experience. Moreover, heterophenomenology also seems to sidestep a possible flaw in classical phenomenology, specifically that it may be so that we can only ever deal with the manifest illusion of our own consciousness thus meaning that the functional scientific reality that underpins it is forever obscured by the 'magic' it produces. If we "[insist] on the third person point of view" (Dennett. 1991a., p. 72)—as heterophenomenology does, then two things happen. Firstly, we gain access to the whole image of conscious experience: we gain epistemic access to the manifest image (as with classical phenomenology) but we *also* gain access to the scientific image on which it depends. Secondly, we "never abandon the

methodological scruples of science" (ibid) and thus we can never be accused of deviating from the "neutrality … [of] objective physical science" (ibid).

2.3 ACHIEVING NEUTRALITY: SELF-DESTRUCTION AND DETACHMENT

Heterophenomenology's respect of the neutrality of physical science and its commitment to replicating the methodology of scientific investigation firmly establishes the method's status as an incarnation of the belief that the eradication of subjectivity, and along with it, the removal of all anthropocentrism—the idea that all investigations have humans as their nucleus—is a hallmark of truth. The assertion that anthropocentrism and subjectivity should be excised from our investigations if they are to be of any serious epistemic worth has long been the darling of philosophical investigation—natural or otherwise. From Plato and Aristotle to the present day, the notion has endured (Dreyfus, 1991., p. 45ff) and, whether actively or passively, shaped the dominant methods of humanity's investigative endeavours. The claim that eradicating subjectivity is the path to truth also enjoys some intuitive plausibility—if there is nothing present to warp our observations, no biases or no errors, then it makes sense to claim that our observations capture reality in a veracious manner. But speaking of objective neutrality's long history or intuitive plausibility tells us nothing about how one acquires this detached neutrality to get in contact with the truth. In general scientific practice, achieving neutrality is little more than guarding against human error or instrumental bias, but this general sense does not properly capture the nature of neutral detachment; if we are to understand how we are to be detached, we must first explore how we are *attached*.

In one's everyday life, one goes about their business, whatever it may be, paying little attention to the intricacies demanded by the actions they perform. One gets up in the morning and brushes one's teeth. Under normal circumstances, this is done in the absence

of explicit thought, and this also holds true for many of our habitual actions. Automaticity, the hallmark of our everyday experience, is afforded to us by the familiarity of our environment and our actions and the mastery engendered by this familiarity. We know how to cope with the world around us and thus we become absorbed by it; in our everyday existence, *we are attached to the world*, we "dwell" (Heidegger, 1927., p. 54ff) in "the world … [that is] familiar in such and such a way" (ibid). This everyday automaticity continues indefinitely, until one experiences a breakdown or disturbance in their ability to perform their actions that disrupts their ability to cope. Say one comes to make their morning coffee, flicks the kettle's switch only to find it non-functional. In an attempt to fix it, they might pick it up to reseat it firmly, try again but to no avail—the kettle is broken. At this point, the agent, will step back and *think*. This "refraining from every manipulation and use" (ibid., p. 61), typifies the detachment spoken of above. As such, we can think of the destruction of the subjective anthropocentric viewpoint to be the endpoint of a process of "withholding of the practical attitude" (Dreyfus, 1991., p. 79), or alternatively, that which "is left over after the cessation of practical activity" (ibid., p. 79).

2.4 ON THE NECESSITY OF SUBJECTIVITY AND CARE

Now that we understand detachment, we can move on to understanding the flaw of the project of attempting to remove anthropocentrism and subjectivity from our investigations: namely, that doing so is impossible. There are two things that we can say about the above formulation of detachment, two things that are so deeply interlinked that it is erroneous to treat them as wholly separate issues—rather they are two aspects of the same phenomenon. The first is that there is a tension between the withholding of the practical attitude and the temporary destruction of a subjective human viewpoint: anthropocentrism is maintained if we formulate detachment as a form of non-practical theorising, for we cannot escape the fact that it is a *human* that is doing the non-practical theorising. This seems to be a flaw in

the formulation of detachment as non-practicality—if a human is there, then so naturally is subjectivity. This tension is the product of the fact that humans are necessary for the investigative situation: because inanimate objects lack perceptive capacities, it follows that the *total* eradication of any anthropocentrism from the investigative situation is impossible without derailing the investigation itself. Put simply, one cannot investigate something that they have no empirical access to.[7]

As such, there must always be some residual subjectivity to all investigative situations. Even in investigative situations of maximal detachment, where one is simply staring at and thinking about the milieu around them, subjectivity creeps in through the way in which the entities of the milieu strike us. As humans, we are shaped by our societies and we are always in the midst of some project or another, and this acculturation and project-orientated way of living imparts subjectivity to the situation by making some things or actions significant or plausible and others insignificant and implausible. As Richard Rorty puts it, "our [subjectivity] is what makes certain options live, or momentous, or forced, while leaving others dead, or trivial, or optional" (Rorty, 1991., p. 13). This notion clearly features in our lived experience. In our fast-paced individualistic Western societies, when our kettles stop working in the morning, we see the possible option of taking it around to our neighbours to get their perspective on how to revive it as a *dead* option— *I don't have time for this, I'll just grab a coffee from the train station and perhaps buy a new kettle tonight*—yet in another, less demanding and more community-driven society the option of seeking your neighbour's expertise may be the most sensible option. As such, subjectivity cannot be eradicated totally so long as the human observer is present. It can only be

[7] Consider the famous thought experiment: if a tree falls in the forest and there is nobody there to hear it, does it make a sound? It operates utilising the same principle: if the investigative agent is not present, then no sound in the sense of an auditory experience is produced, even though sound in the sense of the production of soundwaves is present.

minimised through forgoing practicality: there is “no skyhook provided by some contemporary or yet-to-be-developed science [that] is going to free us from [our subjectivity]” (ibid).

This leads us to our second point: minimising subjectivity is an *active* process. Refraining from manipulation or use is something that is *done*. This suggests that detachment is the negative mode of another phenomenon: investigations that seek to “determin[e] by observation what is objectively present … must first [have] a *deficiency* of having to do with the world” (Heidegger, 1927., p. 61). This evidences a *positive mode* of the phenomenon, a *having to do with the world* that is *not* deficient. If we think of this *deficient* mode of *having to do with the world* as non-practical theorisation, then it follows that the *sufficient* mode of the phenomenon is the inverse of this: practical non-theorisation. Here, we arrive at the automaticity that characterises everyday coping with the world—our autopilot that knows how to squeeze toothpaste from the tube and use door handles without giving them any thought. With these two modes of the phenomenon expressed, we should look more closely at the notion *having to do with the world*, or, to call it the name given by its instigator, Martin Heidegger: Care (*Sorge* in its original German).

Before we get too deep into the notion of Care it might be prudent to clarify just what it is Heidegger means by the term ‘world’—if we do not properly understand what we mean by the term then any discussion of the *having to do with the world* that typifies Care is hindered by ambiguity. Proper dissection of the term reveals that there are four senses of the term ‘world’:

1. ‘World’ as the “totality of beings that can be objectively present in the world” (Heidegger, 1927., p. 64). This sense is the one most readily associated with philosophical and scientific discussions of the world; the term here refers directly to the total aggregate of objects that we may find ourselves amongst.

2. ‘World’ as a term to describe *the being* of the totality of objective beings present in the world—that which these beings have that makes them of the type that they are instead of some other type (Mulhall, 2005., p. 46). For example, the range of equipment used by a maths teacher (rulers, protractors, calculators, etc.) are mathematical and pedagogic by virtue of the fact that they belong to the ‘world’ of maths and teaching, as opposed to some other world (op cit., 1927., p. 64).

3. ‘World’ can also refer to our particular and immediate arena of experience, the *where* at which we are (Mulhall, 2005., p. 47). For example, the offices in which the accountant does their work constitutes a different world to the kitchens of a professional chef, or the comfort of their armchairs at home. This sense is complicated somewhat by the fact that these worlds can be public or private—the accountant’s office, where they sit in an open-plan configuration alongside their colleagues constitutes a public world as it is shared by all those present. Should the accountant then choose to spend their weekends alone, then this world is private, by virtue of the fact of the accountant’s solitude (op cit., 1927., p. 65).

4. The final sense of the term ‘World’ refers directly to the ‘worldliness’ of our immediate arenas of experience. It is “the structural totality of ‘worlds,’ but contains within itself the *a priori* of worldliness in general” (ibid., p. 65)—in

> essence, it is that which makes it possible to think of the cold milieu of *things* that constitutes our immediate arena of experience—our *where*—as 'worldly'.

With this done, we can articulate the significance of Care. As we have already suggested, Care comes in two modes: one negative and deficient, typified by non-practical theorisation, and another, positive and sufficient, typified by non-theoretical practicality. These two modes of Care rest upon the notion of *being-in-the-world*. At the moment of our birth we are thrust into a particular situation—we are "thrown … into [our] there" (ibid., p. 131), our worlds, in the third sense articulated above. This notion of *thrownness (Geworfenheit*), is intimately linked with our worldliness, for it is the world—again, in the third sense above—that we are thrown into. By being born, we are cast into a situation, a particular place that engenders particular practices. Thrownness, in this way, is the requisite of our being-in-the-world; the condition of 'being thrown' is that we are 'in-the-world'.[8]

This notion of 'being-in' is ambiguous: we can explore this ambiguity through semantics. Consider the sentence "Leonid Brezhnev is in the Kremlin". We can think of Brezhnev's 'in-ness' in two senses. The first sense is one that holds the above sentence as representing a fact about Brezhnev's geographical location—he is literally *in* the complex of government buildings that stand in the heart of Moscow. In this sense, 'being-in' captures a fact about location, or of proximity—in this sense, we are *in* the world in the same way as two objectively present things can be in one another. To claim that we are *in the world* in this way reduces our relationship with the world to the same kind of relationship that a

[8] This implies some form of causality—that our 'being-in' is the *product* of our thrownness—which in turn suggests that our 'being-in' is contingent, that some other phenomenon could be the product of our thrownness. This is not so. If we were not thrown into the world, then we would not be able to 'be-in' it.

broom may have with a cupboard—we are in the world "as water is "in" the glass" (ibid., p. 54). This does not capture our usual way of 'being-in'. Consider again the above sentence. The second sense in which we can think of Brezhnev being *in* the Kremlin says less about *where* Brezhnev is and more about *what he is doing*: "Brezhnev is in the Kremlin" can mean *Brezhnev is the General Secretary of the CPSU and is performing the tasks demanded by his role*. This second *active* sense of 'in', which, unlike the first sense does not infer *inclusion* but instead expresses *involvement*, is the sense of 'in' most closely associated with the 'being-in' of a *thrown* being (Dreyfus, 1991., p. 42ff).

This sense of involvement, this sense of *being-engaged-with*, is that to which Care refers. Regardless of whether one Cares in the sufficient practical mode, or the deficient theoretical mode, one is still involved with the world in some way: we are still "*occupied with things*" (ibid., p. 43). We are just as engaged with the world should we step back from our broken kettle as we would be if we immediately picked up our screwdrivers and dismantled it—"the *deficient* modes of omitting, neglecting, renouncing, resting, are also ways of taking care of something" (Heidegger, 1927., p. 57). Our existence has the character of thrownness, which is co-occurrent with our 'being-in', which in turn is the condition of our Care, our tendency to become *involved* with the world.

2.5 ON THE *TELOS* OF CARE AND THE NATURE OF SIGNIFICANCE

Understanding what Care is does not tell us its purpose. This is straightforward— through Care humanity "takes a stand on itself" (Dreyfus, 1991., p. 43). Humans are unique within the world as we are the only beings to which our being is a concern. We are the only beings who Care about ourselves, or, to put it into other words, we are the only beings to ask the question, *what am I?* As such, we are the only beings that are self-interpreting (ibid., p. 23 – 25)—because our essence "lies in [our] existence" (Heidegger, 1927., p. 41),

what our essence *is* "must be understood in terms of [our] being" (ibid). Or, to put things in different terms, what we *are* is determined by whatever "possible way for [us] to be" (ibid) we are at the time that we ask the question '*what am I?*' What we are at any given time is determined by "what [we] do, need, expect, [have] charge of in the things … [which we] *take care of* in the surrounding world" (ibid., p. 116). If we are involved with the chisels, the planes and the saws that we find in our workshop, then we are a carpenter. If we take Care of the powdered wig, the cloak and the case files in our judicial chambers, then we are a barrister.

As we can see, what a person *is*, is based upon the objects that they come to Care about; the objects that they come to circumspect as meaningful, as useful tools for the fulfilment of one's goals. It is a nonsense to say that someone who does not Care for a tennis racquet—someone who does not research the best way to tighten its strings, rewrap its grip-tape and take it up into one's hand and *use* it, for example—is a tennis player, for the Care of a certain set objects is the condition by which we *become* whatever it is typified by those objects. Therefore, it becomes clear that these meaningful things that we Care for sit as nodes within a *network of significance*, and depending on the particular configuration of this network—i.e., what we think of as significant, or not—we *become* different things. The network of significance for a city banker will be different enough from that of a Welsh miner, as to account for the difference between them.

This notion of a network of significance is not able to be grasped in its entirety until one understands the genesis of significance itself. It is obvious that something can only be installed in such a network by the virtue of its significance but this prompts the question of how significance arises. Is there a full and complete network of significance already

present, readymade and poised for discovery, or, is a network of significance something that has to be made, created in the throes of our thrownness? To put the question differently, is a network of significance a *given*, or something that humanity, collectively and individually, has to work to make? It certainly seems that when we are thrown into the world, we are thrown also into some pre-existent network of significance (not every tennis player has to work out that a racquet is significant for themselves, for instance) but this reveals nothing of the ultimate origin of the network itself. Our question regarding the givenness of significance still stands.

Should we take the passive theoretical approach that typifies the deficient mode of Care, it seems to us that the network is indeed something that is lying in wait, poised to be discovered. From this stance, it seems to be that significance (or, insignificance) is a kind of *property* that an object may, or may not have, and that is discoverable—knowable—through investigations that are ultimately empirical in nature: understanding something as significant therefore can be thought of as a type of knowledge about the way the object *is*. By way of example, let us borrow a scenario from Husserl (1931., pp. 39 – 41). Say one walks into a room and before them, they see a die upon a table. This die is unfamiliar to them; it is the first one they have ever seen. Eager to understand the object, they walk slowly around it, observing the die in its entirety. Each step brings with it another slightly different visual image, the catalogue of which is then unified into a singular mental entity; they become one through "*synthesis,* a mode of combination exclusively peculiar to consciousness" (ibid., p.39). Once this synthesis occurs, the solidity of the die is revealed to them alongside its usefulness (playing games of chance, for example) and its position in the network of significance is discovered. Following this example, it is reasonable to suggest there exists a parallel between this kind of 'significance-discovery' and the kind of

empiricism associated with John Locke and his followers. Husserl's notion of combining together multiple visual images into one cohesive mental representation bears a strong resemblance to the distinction between complex ideas and simple ideas found with Lockean empiricism (with the overall significance of Husserl's die taking the form of a complex idea, made up of some kind of simple proto-significant ideas given in each visual frame)—and thus, it seems that arguments that seek to undermine empiricism will also act to undermine the supposed givenness of significance.

2.6 AGAINST THE GIVENNESS OF SIGNIFICANCE

To argue against the understanding of significance as *lying-in-wait*, we shall employ the arguments of two philosophers that "invoke the same argument … against [the distinction between the] given-versus-nongiven" (Rorty, 1979., p. 170)—Wilfred Sellars' argument against the Myth of the Given and W. V. O. Quine's argument against the Two Dogma of Empiricism. We shall begin with the argument against 'givenness' as articulated by Sellars.

Let us return to Husserl's example of discovering the significance of a die through theoretical observation. When the agent observes the die, they are in receipt of sensory information from the die: each frame is characterised by differences in this data. As such, in this instance, 'observation' is treated as being broadly synonymous with 'sensing'. We can assume that this act of sensing is enabled by the possession of sensory organs—eyes, ears, noses, etc.—that receive sensory data in much the same way as a radio receives radio-waves: the sensory data is already present, so access is a matter of possessing the correct equipment. Thus, we can assert that to sense something is to be in receipt of some manner of sense-datum—e.g., when we look at the die, we receive sense-data regarding its colour. The empiricists would thus argue that the specific colour of the die is transmitted *as* the

sense-datum we receive—we know the colour of the dice because "sensing *is* a form of knowing [as] it is *facts* rather than *particulars* which are sensed" (Sellars, 1997., p. 16). Conversely, Sellars' argument holds that the *content* of our sensations—i.e., the sense-datum received—carries no cognitively useful information as such information must be conceptualised inferentially before we can understand its cognitive significance: the brute receipt of a sense-datum cannot tell us *what* (if any) cognitively useful information it carries, because we have to work this out in relation to our other sensations, via language. For givenness to be persuasive, it requires that consciousness alone, with "no learning, no forming of associations [and] no setting up of stimulus-response connections" (ibid., p. 20), is enough for us to find things out about our world. Yet, in Sellars' view, this cannot be done because sensory data is not self-conceptualising: it carries no *facts* about the relevant object. Having a sensation carries no meaning because it has yet to be conceptualised via language—the brute receipt of sense-datum carries no cognitively useful information, because is it *we* that give *it* significance, not *it* that gives its significance to *us*.[9]

Though empiricism is Sellars' intended target, his argument has consequences elsewhere. By attacking givenness, Sellars demonstrates that knowledge is without a foundation—it is not based on the synthesis of atomistic units, as Husserl and Locke would suggest, but rather it is a linguistic endeavour. There is no guarantor to our knowledge, nor any foundation on which it can be built because the objective *way-it-is* of the world is epistemically unavailable: all we have to justify an epistemic claim, according to Sellars, is language. The ramifications of this suggest that there are no elements of our knowledge

[9] Here we can see the influence that Kant had on Sellars. This insight was first Kant's: "neither concepts without an intuition in some way corresponding to them, nor intuition without concepts can yield knowledge" (Kant, 1787., p. 85).

that are *not* the product of linguistic justification—we are given no knowledge for free as all knowledge is the product of our own investigative labour. In such a case, the very notion of givenness is attacked by Sellars—it is not just empirical givens that are undermined, but *all* givens. Any knowledge or argument that relies on givens—i.e., that is *not* the product of our (individual or collective) investigative labour—is undermined by Sellars. Insofar as this holds, Sellars also provides a compelling argument against nativism and all theories that incorporate innate ideas (themselves givens) in their premises. For example, Leibniz' assertion that "truths [are] in us as the figure of Hercules is in the marble" (Leibniz, 1765., p. 46)—i.e., that we have truths sitting innate within us, ready to appear should the right conditions be met—or, more recently, Chomsky's suggestion that our ability to learn language is due to our possession of an "innate language-acquisition system" (Chomsky, 1965., p. 53) are both placed in jeopardy by Sellars' claim that no knowledge is given to us for free.

Sellars is not alone in reaching this conclusion, it was also reached by Quine. The motive behind Quine's *Two Dogma of Empiricism* is to highlight that the distinction employed by the logical positivists between analytic truths—i.e., those that can be known through pure reason alone, "by virtue of meanings" (Quine, 1951., p. 21)—and synthetic truths—i.e., those that can only be known through experience—is incorrect. 'All bachelors are unmarried men', for example, is an analytic truth because the attribute of 'being unmarried' is "already conceptually contained within" (ibid., p. 20) the notion of bachelorhood. On the other hand, 'Howard works at the zoo' is a synthetic truth (if it is true at all) because we would need to *experience* Howard working at the zoo to ratify it—the state of 'working at the zoo' is not conceptually contained within the concept of Howard. Let us take the above analytic statement as an example to work with. We can see

that the two major concepts present in the statement—bachelorhood and unmarried men—are co-referring terms, which entails that within an extensional context they are intersubstitutable *salva veritate*. 'Nestor is an unmarried man' and 'Nestor is a bachelor' both refer to Nestor's sex and Nestor's marital status. Because of this equivalence of reference, the two concepts are synonymous: they *mean* exactly the same thing. Now, let us say that over the course of time the concept of bachelorhood changes. Maybe some religious autocrat comes to rule a country and decrees, on the back of some questionable exegesis, that marriage is not a relationship between two people, but is instead a relationship between a person and God, and declares that all citizens are legally married to God by default in order to guarantee their salvation. In this case, the concept of marriage could shift to refer this legal relationship with God whereas bachelorhood could refer to whether one is or is not currently in a romantic relationship with another person. At this point, the terms 'bachelor' and 'unmarried man' cease to be intersubstitutable—Nestor could be both married and a bachelor at the same time (i.e., he is not currently in a romantic relationship, but is married to God)— and the analytic statement 'all bachelors are unmarried men' breaks down via experience. This erodes another of the supposed foundations to our knowledge, the elimination of metaphysics through the matching up of sensory experiences to linguistic statements analytically. With the notion of analyticity eroded, Quine draws the conclusion that "the totality of our so-called knowledge … from … history to … pure mathematics and logic, is a manmade fabric which impinges a on experience only along the edges" (ibid., p. 39); or to put differently, nothing is given to us from which we can base our epistemology—not logical laws or anything otherwise. Thus all knowledge and all beliefs exist within a totality of inferential and contingent relationships between each other.

Following these two arguments, it becomes clear that the notion that there is an objective given network of significance lying in wait for us to uncover is problematic—because there are no givens, significance cannot be given to humans by the world, rather the world is given significance by humans. This prompts the question of how we can produce a network of significance. Should one follow in Rorty's (1979) footsteps and take the arguments of Sellars and Quine to their logical conclusions, one will find that the installation of a network of significance is a linguistic task (Rorty, 1979., p. 175ff)—what belongs where in the network is based upon where we can *justify* it to be, through language. In essence, this reduces significance down into the simple maxim: *something is significant if we can justify its significance by talking about it*.

This maxim seems to miss out some nuance on the nature of language. Should we look at language, we shall see that one of the dominant functions of any language is rooted in *communicating our interpretations of our situations* (Jakobson, 1960., p. 353ff). Nouns communicate what we think there is in the world, verbs communicate what we think those things are doing, adjectives communicate what we think these things are like, prepositions communicate where we think things are, spatially ('the cat is in the hat') or chronologically ('we went home after we left the supermarket'), and so on. Insofar as language functions in this way—as a *description* of an interpretation of the world—then we can assert that language, or more accurately, the act of talking or writing, is a way of being *involved* with the world, thus designating it as a way of *Caring* about the world that sits astride the two modes of Care—theorisation and practicality. It sits within the theoretical mode by way of its concern with concepts and assignation of terms to interpretations and it sits within the practical mode by way of the fact that language is a tool that must be approached practically to be realised. Speech, writing, and other linguistic behaviours require some

level of practical know-how; we cannot speak if we do not know how to form words with our mouths and we cannot write if we do not know how to hold a pen. If this is true, and the use of language does fit into the two modes of Care, then it follows that Rorty's claim that significance-attribution is a linguistic task is true, but only insofar as it belongs to the structures of Care, thus suggesting that Care—in both of its modes—is the ultimate tool for creating our networks of significance.

2.7 ON THE CONSTRUCTION OF SIGNIFICANCE

If the argument above, i.e., that Care is responsible for creating networks of meaning, is to hold any water, we must do two things. Firstly, we must give some explication of how the two modes of Care (non-theoretical practice and non-practical theorisation) come to imbue our worlds with significance and secondly, we must explore the arguments against such a position. We shall approach these tasks in the above order.

Let us suppose that we have applied to a gameshow and have been accepted. We take a train down to the studios and we find ourselves locked in a room, being filmed, with some sort of challenge before us. We have been given a number of dense foam shapes from which we must assemble a climbable structure that allows us to access a small key dangling from a hook in the ceiling—the key we need in order to escape the room. The successful completion of this task requires us to attribute the dense foam shapes before us with the right meaning. To get the key we need to embed the shapes in the correct place in our network of significance; we need to give each shape a 'this shape goes *here*' kind of significance if we want to escape the room. There are two ways in which we can come to complete our task, depending on which mode of Care we choose to employ. If we employ non-theoretical practice, we can simply grab the shapes and, learning-by-doing, throw them together in any-which-way that allows us to climb them and reach the key, thereby

assigning the shapes their 'this shape goes *here*' significance by practical means. Through actually setting out and building a scalable tower we give the shapes their significance: the largest piece is the base, the L-shaped piece is useful for adding rigidity to the structure, etc. On the other hand, we could also complete our task by standing back from the situation and observing the pieces before us, thinking about how they may fit together, how would be best to assemble them to ensure their stability, and so on. Then, once this is done, we can assemble the pieces, scale them and get our key. In both instances, the significance of the shapes before us, i.e., where they should go, is revealed to us by *being concerned* about the shapes. By Caring about them and becoming *involved* with them, either practically or theoretically, we come to *interpret* them.

The above example allows us to explicate two conditions that must be in place for our interpretive faculties to operate, or, for our Care to give them their significance. The first is that there needs to be more than *one* object present for our interpretive facilities to function—in much the same way that the reception of a single brute sense-datum is meaningless in and of itself, an object on its own has no significance. This is because "there "is" no such thing as *a* useful thing … [because it has to be part of] a totality of useful things [if] this useful thing [is to] be what it is" (Heidegger, 1927., p. 68)—a pen's significance as a writing tool is lost unless one also has ink, paper, a surface to lean against, etc. Without these things accompanying it, the pen is just *there*, existing without significance. The second condition is that we need to have some manner of goal or project underway, for it is in the light of this project that Care gives significance—if we were thrust into the room mentioned above *without* the goal of unlocking the door by building a scalable structure and collecting a key from the ceiling, it is unlikely that the shapes before us would ever be interpreted as 'useful for tower building'. Similarly, the key dangling

from the ceiling would not be interpreted as significant beyond perhaps a token recognition of its novelty. It is from this contextual background of objects and goals that allows our capacities for Care to interpret the world as significant and in what way.

2.8 THE INTERPRETED WORLD: CORRELATIONISM VS NEOREALISM

The claim that the world is always interpreted is a radical claim and thus enjoys significant opposition. The emergence of contemporary *post-phenomenological* philosophical movements such as speculative realism and the associated new realism[10] have prompted a re-emergence of metaphysical realism. These neorealist movements arose as a reaction against phenomenology and the supposedly anti-realist conclusions that phenomenology seems to support, i.e., "that subjectivity and objectivity cannot be understood or analysed apart from one another because both are always already intertwined or internally related" (Zahavi, 2016., p. 293). This common enemy of neorealism, i.e., the erosion of objectivity and givenness, is known as correlationism, by virtue of the fact that we only have access to the "correlation between thinking (theory) and being (reality) [thus meaning that we] cannot get outside [of ourselves] in order to compare the world as it is 'in itself' with the world as it is 'for us'" (ibid). As we can see correlationism has considerable resonance with the Rortyan notion that we lack a "skyhook" (Rorty, 1991., p. 13) to pull us up and out of our own contextual existence in order to eradicate our subjectivity and see the world as it actually is. If the neorealist critiques of correlationism are persuasive, they threaten to undermine the argument that networks of significance are not given but are rather constructed via interpretation. Neorealism comes in two main flavours—Graham Harman's speculative realism and Maurizio Ferraris' new realism: both proffer different arguments against correlationism. We shall begin by looking at Harman's speculative realism.

[10] New realism, in this instance, refers to the views typified by Maurizio Ferraris (2014), *not* the views associated with the new realism found in 20th century USA.

It is argued by Harman that the split between objectivity—the world as it is, in and of itself—and subjectivity—the way it is to us—is a dichotomy than can be traced back to the Kantian distinction between phenomena and noumena (Zahavi, 2016., p. 293) and that the only way that we can break free of the anti-reality of correlationism is to "[reject] the post-Kantian obsession with a single relational gap between people and objects" (Harman, 2011., p. 5). This rejection is to be done by denying that there is anything particularly privileged or special about the relationship that humans enjoy with the world, thereby advancing the position that "all relations in the cosmos, whether it be the perceptual clearing between humans and world, the corrosive effect of acid on lime stone, or a slap-fight between orangutans in Borneo, are on precisely the same philosophical footing" (Harman, 2005., p. 75). Dissolving the distinction between subject and object (such as a human has with the world) and object and object (such as an apple has with a fruit bowl) relations can be done in two ways: either by levelling the subject down into the objective, or by elevating the object to the subjective—the latter is employed by Harman. The elevation of the object to the subject suggests that there is no significant way in which objects are different from humans and this naturally carries the implication that objects are replete with subjectivity, just like us. Therefore, the closing of the gap between the phenomena and the noumena (at least, in the way that Harman proposes) requires an endorsement of panpsychism (Harman, 2011., p. 120ff)—a controversial position, but one that may be worth it if it allows us to close the gap between the phenomenal and the noumenal.

Yet, as Zahavi (2016., pp. 295 – 296) highlights, speculative realism fails to close this gap. Harman argues that science is also guilty of correlationism and that scientific theories of 'what-this-is' are ultimately interpretations too: "there is nothing the least bit independent

about objects as defined by the natural sciences" (Harman, 2011., p. 53). This prompts the question of how we are to properly access reality. If phenomenology and science are both kinds of correlationism, then it is not clear how we can access reality in any meaningful way. Harman believes that we cannot. The true nature of things, to Harman, is necessarily inaccessible; we can only ever access the appearance of a thing. We can see this in the various places that he talks about how "[objects are] concealed unit[s] … that emit sensual qualities into the phenomenal sphere" (ibid., p. 98), or how "*real* [objects] withdraw into subterranean depths *beyond all access*" (ibid., p. 77; second emphasis mine).Yet, by arguing that the ultimate nature of things is inaccessible, Harman seems to be articulating a position that is largely the same as that which he is supposedly criticising: "Harman criticises phenomenology for its alleged anti-realism and argues that it chains us to the phenomenal … whatever merit there is to this criticism, it certainly seems like a rather fitting description of his own position" (op cit., 2016., p. 296). As such, Harman's variety of neorealism can be dismissed on the grounds that it is itself vulnerable to the very same criticisms Harman levels at phenomenology.

With Harman's version of neorealism dismissed, we can turn our attention to the new realism espoused by Ferraris. Motivated by the same kind of anti-correlationist sentiment (or, anti-constructionist, as Ferraris prefers to call it) as Harman, Ferraris seeks to argue against what he sees as the constructionist's position: that the world is ultimately "constructed by conceptual schemes and … is therefore amorphous and indeterminate in itself" (Ferraris, 2014., p. 31), a position supported by the purported fact that "the constructionist … identifies being and knowledge" (ibid). The ends of such a position, according to Ferraris, are to achieve "wonder, [through] the removal of the obvious [and] the formulation of *fashionable nonsense*" (ibid). Ferraris seeks to guard against the

formulation of fashionable nonsense by directly arguing against what he sees as the most fundamental constructionist claim—the lack of distinction between being and knowledge: "the realist does not merely say that reality exists … [they also support] a thesis that the constructionists deny, namely, that it is not true that being and knowing are the same" (ibid., p. 32). To demonstrate the distinction between knowing (epistemology) and being (ontology) Ferraris articulates a number of characteristics that can be used to demonstrate the difference between the two: knowing is historical, rooted in language, teleological, associated with the 'interior' and is manifestly amendable, whereas being possesses the precise inverse of all of these characteristics: i.e., is a-linguistic, ahistorical, 'external', and so on (ibid., p. 34).

Of these differentiating characteristics, it is perhaps the notion of *amendability* that causes the most upset for the supposedly constructionist conclusions of phenomenology. The notion is simple: its basic claim is that reality—being—offers resistance to interpretation and thus is unamendable, whereas knowledge yields easily to interpretation: "I may or may not know that water is H_2O; I will get wet anyway, and I will not be able to dry up by means of the thought that hydrogen and oxygen as such are not wet" (ibid). One cannot deny that this argument possesses intuitive plausibility. Try as they might, we cannot *interpret* a hammer into becoming a teapot, nor is it possible to argue that "the handle of the coffee pot [becoming] hot if we leave it on the fire" (ibid., p. 35) is the manifestation of some interpretive process.

Yet, wielding this notion of (un)ammendability against constructionism betrays the fact that Ferraris is operating with a fundamentally uncharitable interpretation of phenomenology. It is unlikely that any serious proponent of phenomenology would claim

that we can think ourselves dry, as Ferraris suggests. Instead, the claim of the constructionist is much less pointed than the one presented by Ferraris: it is simply that the objective character of our immediate environments—the way that our surroundings are—is, *in and of itself*, meaningless and uninteresting *until we give them meaning*. For humans to be *interested* in their surroundings, they must undergo some kind of conceptual analysis and be incorporated into a network of significance. It is their place in this network that makes them *of interest*—until this happens, our surroundings are just *mere things*, and we are Sartre's Roquentin standing aghast before his tree (Sartre, 1938., p. 182ff). The phenomenologist does not argue that we can *think* a wooden chair into becoming a cotton shirt, they simply argue that without the notion of 'chairness' or 'woodenness' and without the ability to usefully contrast it against the cotton shirt, the chair is just an insignificant, weirdly-shaped chunk of *something*.

This is the ultimate failure of new realism: an implicit faith in the myth of the given haunts the theory, prompting Ferraris' uncharitable reading of phenomenology. Ferraris, like the empiricists before him, fails to understand that everything must first flow through our subjectivity before it can have meaning. That what happens when something falls into molecules of hydrogen and oxygen is not 'getting wet' if there is not the subjective investigative agent there to interpret the change from one state to the other as such. The phenomenologist does not deny that there is a change of state taking place—this would be absurd. Rather they are saying that such a change in state is *meaningless* in and of itself, in the same way that the pure sensations found in Sellars' work are without any epistemic significance before they are conceptualised. 'Wetness' is not something *given* to us by the world, it is a product of our interpretive endeavours. To suggest, as Ferraris does, that the phenomenologist thinks that we are capable of using our subjectivity to bend the objective

to our wills is to erect a strawman. Phenomenology does not conflate knowing and being, rather it recognises the close and symbiotic relationship between epistemology and ontology by noting that everything worth talking about is the product of conceptualisation and interpretation. Without interpreting and incorporating that which we encounter into a network of significance, the *world as significant* is lost to us; our awareness is reduced to that of a CCTV camera trained onto the main room of a nightclub that knows only flashes of coloured light, profoundly oblivious to the rich bustle of emotion, humanity and significance occurring beneath its gaze.

2.9 HEIDEGGER: INTERPRETATION AND DISTORTION

With the arguments of the neorealists shown to be problematic or unpersuasive, we are free to return to our assertion that the *world as significant* shows up to us as the fruit of interpretative labour. At birth, we are thrown into a world full of *things* that we must then conceptualise and interpret, installing them, through Care, into a network of significance—a revisable conceptual storehouse which we then use to help us to achieve our goals. Through this process of Care, we interpret the world and give it significance. To understand the importance of significance-attribution in the context of our argument against IMF, we must shape the remaining pieces of the puzzle—we must articulate the Heideggerian distinction between the ontic and the ontological before then introducing and explaining Heidegger's notion of *aletheia* (truth).

Imagine a person who has recently inherited the estate of an estranged and eccentric relative. They come to collect their new property only to be greeted by a house, filled with a collection of unusual bric-a-brac as diverse as it is large. The heir is charged with sorting the unusual objects into various categories, yet they have received no guidance as to what these categories should be—the groupings are wholly at the mercy of the heir. To begin,

the heir has to first establish what the unusual objects before them *are*. They have to take the entity in question and work out what makes it *the kind of thing that it is* rather than a thing of another kind, or, to put it in Heideggerian terms, they have to engage in a process of *ontological enquiry*. They have to interpret the objects before them as being in possession of a particular mode of being, on the basis of which they can then form their groupings (Mulhall, 2005., p. 6). By way of example, the heir would have to interpret the Royal Doulton figurines as having the precise mode of being of a Royal Doulton figurine—perhaps this would be motivated by looking at the material that the figurine is fashioned from, its form, its fragility, the particular clang it makes when you tap it or so on—and *not* the mode of being that is possessed by the African ceremonial masks that they were found alongside. This process of working out what there is by isolating and contrasting the *what-being* of entities is, as we have said, an ontological endeavour—when we work out what there is in the world, we are doing ontology. Doing ontology is the precursor to being able to engage with what Heidegger calls the ontic, for ontology is fundamental to our understanding of the world: we first need to know what there is before we can come to know anything about it. Insofar as this is true, we can think of engaging in ontology (i.e., interpreting a thing as being the kind of thing it is, rather than another kind of thing) as the process that gives us the "basic concepts" (Heidegger, 1927., p. 9) needed to engage with the ontic: "the totality of beings can, with respect to its various domains, become the field where particular domains of knowledge are exposed and delimited" (ibid., p. 8). Ontic knowledge therefore, refers to particular domains of knowledge that concerns particular sets of beings, which then informs our understanding of the beings in question with regards to everything that is not established on an ontological level.

Let us return to the heir sorting through the house full of bric-a-brac and assume that they have finished their task and before them sits multiple piles of well organised curios—they have established what there is. They now have before them the basic concepts needed to understand those objects in a number of different ways— let us focus on the pile of rare books our heir has isolated. There are a number of ways that the heir can come to understand these books—the heir can establish a number of things about the books, things that belong to a diverse array of domains of ontic knowledge. They can understand their masses, and how they interact with each other in terms of physics. They can come to understand their economic value as collectors' items. They can read them and come to appreciate them as literary objects, studying how the authors craft their prose and so on. These different interpretations of the books (e.g., seeing the book as a commodity, or a work of craftsmanship, etc.) are the ontic characteristics of the books and the facts pertaining to these interpretations are instances of ontic knowledge (op cit., 2005., p. 4).

As we can see, the notion of the ontic is intrinsically linked with disciplines of study: each different discipline (mathematics, history, philosophy, physics, etc.) is concerned with a different set of ontic characteristics—a different interpretation—of the objects delineated through ontological investigation. We can see thus the role that ontology has in shaping the ontic. Each ontic domain of study is built upon ontological presuppositions; every discipline has a set of *things* that belong to it, whether this be a broad and indiscriminate set of things—such as is the case with chemistry, where almost every concretely existing thing can be analysed in terms of its chemical make-up—or a narrow and discerning set of things—such as is the case with literary theory, that is concerned only with a specific kind of text, having no care for anything else. But simultaneously, these sets of things—the ontological presuppositions on which we base our ontic investigations—are also

interpreted through the lens of whatever ontic discipline you are analysing the thing in the name of; a physicist, a chef and a pomologist will, for example, all see the same apple as a very different thing, ontically speaking.

The importance of the above for our ends is keenest when combined with Heidegger's concept of *aletheia*. *Aletheia*, originally posited by Heidegger as a theory of truth (Heidegger, 1954., p. 317ff) that puts us back in touch with the notion as it was originally conceived in Ancient Greece[11] is best thought of as the "bringing-forth … out of concealment into unconcealment" (ibid., p. 317), or, more simply, a process of disclosure. When one interprets a *something*, they are drawing an understanding of the object out from 'concealment'. To take a hammer and set about using it to drive nails into a board, we are disclosing the hammer as *useful for driving nails into boards*. Of course, as the motif of 'bringing into unconcealment' implies, the concept of *aletheia* is somewhat complicated by the fact that unconcealment is the positive mode of a negative phenomenon; one cannot have unconcealment unless one also posits the existence of its inverted twin, concealment (Pattison, 2000., p. 51ff). This suggests that *aletheia* is as much to do with bringing things out from unconcealment as it is to do with pushing other things *into* concealment. This elucidates a common feature of our lived experience that is often overlooked, perhaps by virtue of its ubiquity—enduring presence tends to lead to invisibility. Think of a time one has become engrossed in a novel and as a consequence, the world beyond that which is given to us by the author is temporarily lost to us—the world 'out there' is obscured by the world 'in' the book. As we engage with the book-world, the real-world is pushed into

[11] Heidegger, in his later work, rowed back from the claim that *aletheia* is a theory of truth on the grounds that there is a distinction to be made between the unconcealed and the true, i.e., something can be unconcealed, yet still be untrue: "to raise the question of Aletheia … is not the same as raising the question of truth … it was inadequate and misleading to call Aletheia … truth" (Heidegger, 1969., p. 70). This does not impact on the importance of the notion of *aletheia* in the context of our argument, however.

obscurity. A similar process is happening here now; the internal monologue prompted by these very words obscures their typeface and any thoughts about the paper that they are printed upon. By paying attention to one thing, something else is unavoidably left neglected (ibid., pp. 50 – 52).[12] This process of simultaneous unconcealing and concealing is precisely that which Heidegger seeks to highlight with his concept of *aletheia*. *Aletheia* envisages disclosure as a spotlight; as one interpretation bathes in light, another is cloaked in darkness.

This leads us to the question of *why* one interpretation is disclosed at the expense of another: why does the chef disclose an apple as *useful for baking into a pie*, whilst this is lost to Isaac Newton who discloses the apple as a useful tool for rendering apparent the effects of gravity? Or, more generally, how can the same entity come to possess an array of different ontic characteristics? It is possible to highlight a number of variables that may impact upon that which is disclosed and that which is left concealed, with regards to an entity. Among the most pressing are:

1. **The characteristics of the being in question.** An entity, interpreted ontologically as being what it is, *is thus what it is*—we cannot interpret a horse into having the same kind of being as a garden chair, nor can we interpret a swan as having the being of a handbag.
2. **Our goals.** The same entity can be interpreted differently on the basis of what it is that we want to do; the hammer used to drive nails is interpreted very differently

[12] An enlightening parallel can be drawn here with the notion of Gestalt—as one thing is drawn into the foreground, another is pushed into the background.

should we be in a draughty workshop and we are in need of something to weigh down our plans.

3. **The milieu of beings the entity in question is encountered amongst.** A medieval knight will interpret his sword differently should it fall into the hands of his enemy after he is disarmed in combat. Similarly, a ring found in the gutter is likely to be interpreted as lost or discarded, should the same ring have been encountered in a jeweller's window amongst other similar items, then it will likely be interpreted as being available for purchase.
4. **The *mood* in which we encounter the being.** The mood in which we encounter the being can influence how it is interpreted. A loud bang may not even register as significant when it is encountered in the middle of the day when one is in an industrious mood and amongst others unfazed by the sound, but the same sound may be perceived as threatening or dangerous when it is met in the small hours of the night, when one is alone and apprehensive (Heidegger, 1927., pp. 130 – 138 : Dreyfus, 1991., pp. 168 – 183).

These four factors (and almost certainly others, yet to be unconcealed) influence our interpretation of an encountered being; the *significance* of the being is informed by these variables, but how does this mesh with the notion that Care is the ultimate source of significance? Recall that Care comes in two modes—a theoretical and a practical—and let us highlight that the theoretical mode is derived from the negative modification of the positive attached practical mode of Care: "pure disinterestedness is an abnormal state … [because] detached contemplation [is] a privative modification of everyday involvement" (Dreyfus, 1991., p. 47). By combining these two notions, we see that everyday attachment is influenced by the totality of *all* (relevant) interpretation-impacting variables—i.e., the

significance of that which is interpreted through practical means is the aggregate totality of our moods, our goals, our societal norms, etc.—whereas detached theorisation purposely suppresses as many variables as it can, in order to put us in contact with what is supposedly the true nature of the entity in question. Each unnecessary variable is eliminated for fear that it may be a distorting influence. For example, the milieu of beings that a thing is found amongst is ignored by detached observation; that which we are investigating is bracketed off and isolated away from its surroundings in order for us to get in touch with its purported reality. Attached practical manipulation, on the other hand, uses the milieu to help it to understand what the thing is and what it might be useful for.

Thus the two modes of Care can produce two different interpretations of the same object. To again use a hammer as an example: attached manipulation can find out ontic facts about the hammer that are unable to be grasped by the deficient mode of Care. Say one encounters a hammer for the first time, in complete isolation of the things it is most often associated with (nails, walls, boards, etc.). Its efficacy for driving nails will be concealed should one not take the object into one's hand to feel its weight and understand how one can use it. One cannot circumspect—see the usefulness of—the *what-for* of the hammer on one's first disinterested encounter with the object; that comes afterwards, when the usefulness of the hammer has somehow already been disclosed. All that maximally detached observation can truly disclose about the hammer is that it seems to be extended—it seems to take up a certain amount of space and be in possession of a certain form—whereas the positive mode of Care can reveal ontic knowledge about the hammer's use as a tool, its status as a historical and social object, its status as an object of manufacture, etc.

The importance of this is especially keen should one combine it with the concept of *aletheia*—different modes of Care lead to different interpretations of the same object by way of the interpretation-impacting variables they apply to the entity in question. Yet, these interpretations are *competing* interpretations: when one is interpreting the hammer as *useful for driving nails* one is not interpreting it as an extended object, for instance.[13] This means that the dogmatic adherence to a particular interpretive methodology (e.g., the insistence on a detached *view from nowhere*) naturally causes an inevitable ignorance of all that could be uncovered should we use a different interpretive methodology. If we focus on detachment, we lose the *hammer as tool* or the *hammer as historical* behind the *hammer as extended*. The methodology by which we try to install things as nodes in our networks of significance impacts upon the place that they ultimately reside in them: *how* we interpret an entity has an unavoidable impact on what the entity is interpreted *as*.

2.10 IMF AS THE PRODUCT OF A DISTORTING METHODOLOGY

In §1.3, we isolated IMF as the image of humanity that technology seeks to capture—an image that accounts for mentality in functional terms. This understanding of humanity allows developers of humanlike technologies—in theory—to create them after our own image: all there is to making a humanlike technology is creating a system with human functional parity. In addition, the project of creating humanlike technologies is aided by IMF's illusionist sympathies: not only is the project of creating humanlike technologies simplified somewhat (developers do not need to worry about recreating phenomenal consciousness, as it does not actually exist) but it is also guarded against any detractors that would argue that the inability to materially reproduce consciousness would hinder the

[13] As we come to master the skills needed to use an object, the object becomes *phenomenally transparent*, i.e., we stop noticing that we are manipulating an object. Think about scribbling down a message during a telephone call; the pen you use never presents itself as a brute object because this 'thingliness' is lost, obscured somewhere behind the active use of the object.

project. Put briefly, developers simply do not have to worry about the issue if consciousness is naught but an illusion produced by the functional workings of our brains. These boons to the project of creating humanlike technologies though are only present if IMF itself is justified. If IMF itself is not a viable position, then these boons are lost.

IMF sees the job of philosophers of mind, neuroscientists, cognitive scientists, etc., as that of finding and explaining *how* the illusion of consciousness is produced: "[it is] our burden to explain how the 'magic' is done" (Dennett, 2017., p. 65ff). We have seen that this task of 'explaining the magic' behind the illusion of consciousness is to be achieved, in part, via Dennett's heterophenomenology—a way of collecting data about peoples' beliefs towards their own conscious experience that amounts to asking people what they believe is true about their conscious experiences. This methodology has the purported advantage of allowing us access to both the manifest image (i.e., the other person's supposed conscious experience) and the scientific image (i.e., the other person's neural activity/functions) that supports the illusion, Moreover, it does this whilst "never abandon[ing] the methodological scruples of science" (ibid) or deviating from the "neutrality … [of] objective physical science" (ibid). It is this commitment to the scruples of science that cements IMF as belonging to the philosophical tradition that holds detachment—the (impossible) suppression of subjectivity—to be the gold standard of investigative methodologies. But, as we have seen in this chapter, the deficient mode of care that typifies detachment and disinterest is only able to uncover a *limited* amount of information about the object in question (i.e., its extension, the modes of its extension and the functions thereof). Moreover, this information is revealed at the expense of *other* information about the object, as suggested by *aletheia*: shape, size, location, movement, etc., are revealed at the expense of any other ontic characteristics.

Whilst interpreting worldly entities through a maximally detached mode of Care is a perfectly respectable endeavour in its own right, care should be taken to ensure it is not misapplied: when we want to find out about quantifiable material entities, it is the very best mode of interpretation that we have at our disposal, but to claim it to be universally applicable to *all* ontic investigations is misguided, yet this is what materialism encourages us to do. Misapplying the detached mode of Care undermines IMF in two ways. Firstly, it casts doubt upon the supposed objectivity of Dennett's methodology. It seems that Dennett is taking that which is the product of detached investigations—i.e., extension, and the modes and functions thereof—and *interpreting* them as belonging to the discipline of objective physical science, which itself is interpreted as belonging to metaphysical materialism. It seems *ad hoc* and hypocritical to dismiss one interpretation over another, on the grounds that the competing theories are themselves interpretations and thus cannot be trusted. Secondly, it shrinks down one's explanatory toolkit unjustly, to the point that one can only use the quantifiable terms of material entities to describe and understand our world. For some applications, this toolkit is all one needs; most investigations concerning the way that objects are do not seem to need more than this, but *humans are not objects*. To interpret a human through the deficient mode of Care—to equip oneself with the limited toolkit implied by IMF— engenders a situation where we take up the role of an interpretive agent, sitting in contemplation of an *object* (Dreyfus, 1991., p. 45) in possession of a certain fixed set of properties—a certain way it has to be—that is lying in waiting to be discovered. Doing this though, reveals the way we are *object-like*, at the expense of the way we are *human-like*. It is erroneously presupposing that the ontology of a human is identical to that of an object.

Thus the image of humanity IMF presents is an *impoverished* image: we reduce humanity, in its grand and intricate totality, into mere human bodies and the functions thereof, and so we have no grounds beyond an appeal to the prevailing materialist doxa to assume that IMF properly captures human consciousness. Clearly, for questions concerning anatomy, physiology, etc., this presents no issues, for the target of their investigations is the human body—the *human as object*. But when we begin to ask questions after human consciousness and being, it is not appropriate to treat the human as an object, if for naught else, because we are committing a category error: objects are not conscious and it is consciousness with which we are concerned. To treat humans as if they are objects is to eliminate from the equation the very thing we seek to explain—it is unlikely that any theory of consciousness articulated in terms of the way that humans resemble objects will be satisfactory, simply due to the fact that object-hood is a mode of appearance *within* consciousness. Any object-motivated analysis—i.e., any interpretation of humankind that views us in the same way as objects is doomed to missing out consciousness—unpersuasive or unhelpful treatments of consciousness are the necessary consequence of trying to squash humanity into a category which it does not properly belong.

Insofar as the above holds, it does not seem that IMF actually says much of value about consciousness beyond the assertion that it does not actually exist. Its insistence on holding up the neutrality of objective science forces it into a particular ontic understanding of the world—one that leaves in the dark the very thing it seeks to bring into the light. To focus on the way we resemble objects is to neglect the way we differ from objects, thus leaving the unique peculiarities of humankind obscured. This is a fatal move for a theory of consciousness to make: our consciousness is the most salient way in which we differ from objects. The insistence that detachment is the best way to discover reality binds those

convinced of IMF to their illusionist conclusions—if consciousness is obscured by their ontic interpretation of humans, then it is so that they cannot explain it with the terms given to them by their interpretation. In the same way that one cannot explain the French revolution with algebra, one cannot explain consciousness with metaphysical materialism. Material entities and the modes and functions thereof are not explanatorily sufficient for consciousness, so any theory trying to explain consciousness using them eventually arrives to the conclusion of that consciousness is an illusion promulgated by the workings and functions of the material brain. Yet this is unsatisfactory in the same way that it is unsatisfactory to claim that the mass of an apple is illusory because you cannot measure it with a thermometer—if your theoretical toolkit cannot help you to understand something that is obviously manifest (even those convinced of IMF would claim that it does *seem* that we are conscious—there would be little point in revealing the illusion otherwise) then the proper response is not to cling to your toolkit and posit a *de facto* denial of the existence of that very thing which you are trying to explain, but rather to expand your toolkit to allow you to do what you set out to do. The phenomena reveal that we must adapt our current toolkit to cope with any explanatory shortcomings it may engender—it is unsatisfactory to deface and truncate the world so it can be explained by the few tools you already have. One cannot claim that the fence does not *really* need a fresh coat of creosote and just *looks* as if it does, because all that is in their toolkit is set of screwdrivers—to do so is clearly absurd, yet, this is broadly the move made by IMF.

As such, we can see how IMF comes to present an impoverished image of humanity. It is done on the grounds of a distorting methodology bequeathed to the investigator by the prevailing materialist doxa. By taking the view from nowhere, IMF denies itself the conceptual toolkit needed to talk about consciousness in a meaningful way—human

subjectivity is lost, obscured by a methodology that is most useful for understanding and explaining the objective—and therefore, it is prudent to be wary of the illusionist conclusions that it draws about consciousness. This hampers the development of humanlike technologies in two ways. Firstly, if IMF is the product of a distorting methodology, then it makes sense to argue that the boons it gives to the project of developing humanlike technologies—i.e., that of simplifying the process, and the ability to guard against objections that suggest the project is impossible as consciousness cannot be materially reproduced—are lost. If the image of humanity it presents is not accurate, then developers of humanlike technologies *do* have to worry about reproducing consciousness. Secondly, the project does not just suffer through the loss of boons. If IMF provides the blueprint of humanity that guides the development of humanlike technologies, and its illusionism is not justified as it is a conclusion reached via an inappropriate investigative methodology, then naturally this incurs a penalty to the humanlike-ness of these technologies. We cannot develop them to be humanlike if our understanding of what it is to be a human is impoverished: omissions in the plans are poised to be translated into omissions in the product.

In the following two chapters, we will continue our exploration of how the image of humanity advanced by materialism is impoverished: we shall, in the next chapter, explore its difficulties in accounting for free will, and afterwards, in chapter 4, we shall see how materialism faces similar difficulties in accounting for our sense of selfhood.

3 THE IMPOVERISHED IMAGE II: MATERIALISM, THE CLOSED WORLD AND FREEDOM

Materialism (and by extension, IMF) does not just face issues in trying to explain our consciousness: it also has difficulties in accounting for and explaining the free will that the basic phenomenology of being-human suggests that we have. To demonstrate this, we shall focus on two arguments concerning free will. Helen Steward's *a priori* argument *for* free will—Agency Incompatibilism—will first be considered and shown to be unattractive, before then shifting our attention to Benjamin Libet's well-known *a posteriori* argument *against* free will and its flaws. This is done to highlight that the failure of both of these arguments is owed to a shared feature of both arguments: a misunderstanding of the proper nature of free will that arises from the detached investigations demanded by materialism.

An appropriate gloss on materialism[14] is that all phenomena either supervene upon the material in one way or another, or are themselves material: all phenomena can be accounted for in material terms, either directly—in the case that the phenomenon is itself material—or indirectly—where the phenomenon is somehow necessitated by the material. Furthermore, observation of the material has allowed for the articulation of physical laws that seem to govern the behaviours of matter and the transition from one material state of affairs into another. The existence of physical laws has considerable impact on the notion of free will: as our actions can be accounted for in material terms, and matter operates in accordance with its own set of laws, then it follows that our actions are also governed by these same laws, and thus, what we do is "not up to us" (Van Inwagen, 1983., p. 16). Or, to furnish the same argument in different terms, if "all physical events are caused or

[14] We shall explore what it means to be a materialist in greater depth in chapter 5.

determined by the sum total of all prior events … then our every deed and decision is the inexorable outcome, it seems, of the sum of physical forces acting at the moment, which in turn is the inexorable outcome of the forces acting an instant before, and so on, to the beginning of time" (Dennett, 1984., p. 1). With such an understanding it is clear that our actions, in the context of materialism, are vulnerable to being characterised not as occurrences that emerge from our undetermined volition, but instead as events that have befallen us by virtue of a deterministic causal chain that sits beyond our influence.

One avenue of response that the materialist may make when met with this argument is that it operates with an outdated understanding of physics. Materialism may seek to undermine the suggestion that the material is subject to the laws that govern their behaviours by appealing to the randomness suggested by some non-deterministic understandings of quantum mechanics.[15] The transition from one state to another is not an orderly process as suggested by physical laws, but rather it is one marked by randomness and as such, the crux of the above argument is supposedly undermined, i.e., our actions are not owed to physical laws. This response is flawed as it misunderstands the relevant implication of determinism—that our actions are not a product of our own volition. Replacing the laws of physics with quantum randomness does nothing to reassert our dominion over our own actions; it simply swaps out an orderly master for a chaotic one. In both cases, we are unfree to act in any other way than how we do: "neither of the two options, *determined* or *random*, seems able to give us … what we want" (Strawson, 1986., p. 21).

[15] For instance, the view of quantum mechanics found in the '*chaos and quantum theory*' understanding of free will, found in Kane 2005, p. 133ff.

3.1 STEWARD'S *A PRIORI* CASE FOR COMPATIBILISM

There is a better way of criticising the determinist's argument to try and preserve free will within a materialist framework. Should we return to the argument, we notice that the determinism central to it is totalising—no worldly phenomena is free from determination. This is a strong claim so presents itself as a weak target: should one demonstrate that there is some worldly phenomenon or other that is not determined, then the space needed for free will is secured. A phenomenological approach suggests that the best candidate for this undetermined phenomenon is our own capacity for self-movement—nothing seems to be quite so obviously a product of our own will as when and how we move our bodies. This is a view shared by Helen Steward and motivates her Agency Incompatibilism. Steward argues that the existence of self-moving animals—of which we humans are but one of many kinds—is sufficient grounds to demonstrate "the falsity of universal determinism" (Steward, 2012., p. 12). Steward suggests that the determined world is one that is characterised by its closedness—if the unfolding of events follows a determined sequence, the future is no longer open: "whatever happens anywhere in the universe … is necessitated by prior events and circumstances" (ibid., p. 9). Yet for Steward the existence of self-moving animals has the ability to arrest and subvert the sequence of happenings in the world—"where animal agents exist in a world, the unfolding of that world through time must wait upon decisions and choices which have to be made by those animals" (ibid., p. 18)—thus opening up the future to an array of different realisable potentialities. These ideas can be formalised into a master argument to undermine universal determinism:

1. If universal determinism is true, the future is not open.
2. If there are self-moving animals, the future is open.
3. There are self-moving animals.
4. Therefore, universal determinism is not true. (ibid., p. 12)

Steward's argument runs into an initial difficulty because it suggests that the self-moving capacities of animals *en masse* are sufficient grounds to secure free will the space it needs to exist. Such a category contains within it both very simple organisms and some larger, more complex organisms to which we would not intuitively attribute free will. Consider very simple organisms, such as the paramecium that Steward uses to illustrate this objection. Paramecia are single celled aquatic organisms that propel themselves through water through performing a rapid 'rowing' action with the cilia present on its body. In the pursuit of preferable conditions, the paramecium will propel itself, until it reaches an obstacle, where it then "backs up, turns and progresses forwards until another stimulus is encountered" (ibid., p. 15). It seems difficult to suggest that such an organism is not self-moving, yet it is similarly difficult to suggest that the existence of such an organism is enough to secure the conceptual space needed to argue for the existence of free will. Steward's solution to this is simple—whilst it is undoubtable that the paramecium is self-moving, such self-movement does not constitute the *kind* of self-movement that is key to the argument she presents.

The salient kind of self-movement for Steward is found in larger, more complex organisms where one can feasibly argue for a distinction between "the creature and its body, an entity to which the creature itself stands in the relation of controller and director" (ibid)—a paramecium is not "sufficiently complex to sustain an owner/body distinction" (ibid., p. 16) and thus cannot be counted as an agent. Naturally, this response prompts calls for a deeper exposition of agents and agency—this shall be returned to shortly. But for now, even with this problem addressed, it seems as though there is an issue with attributing free will to more complex organisms—where a distinction between creature and body is plausible—because of the role that instinct plays in the actions of these animals. It does not

seem that a swallow, migrating back and forth between hemispheres with the seasons does so of its own volition—the swallow cannot choose to 'sit out' the migration for a season and remain in place. Steward concedes that animals do not "have any grand capacity to transcend … promptings of instinct" (ibid., p. 20) but denies that this has any great impact on the argument she seeks to make. The locus of free action, she suggests, does not dwell in the thwarting of instinct but rather the capacity to respond to such instinctual constraints in a plurality of non-determined ways. Agency is exhibited not by the swallow sheltering in place, but rather in the free choosing between the number of different possible flight paths available to it when migrating, its precise location in the migratory flock, etc.—the action of migration is indeed unfree, but the *technique* of migrating is under the remit of the swallow itself, and thus is an exercise of its agency.

For Steward, the existence of determinism-defeating agency relies upon the notion of "settling a matter" (ibid., p. 39), or *settling* for short. The idea is that the unfolding of events frequently contains within its course events that are not already resolved by reference to their antecedent events, and thus present themselves as an opportunity for an animal agent to pick between the different potentialities of resolution themselves. Once the potentiality of resolution is decided and acted upon, the course of events continues and the matter is settled. *Settling* in general is ultimately realised by way of the agent effecting change in its own body, either in an outward, overt manner—such as actual bodily movements, e.g., raising one's arm—or in an inward, covert manner—such as the modulation of the microphysical conditions within the agent's own body, e.g., bringing about "the occurrence of certain brain events" (ibid., p. 45). The agent is able to control its own bodily settlings because the agent is—uniquely—in possession of what Steward calls

"two-way powers" (ibid., p. 126). It is through the exercise of these powers that the agent can instigate and oversee changes in their own bodies.

Two-way powers are straightforward—they are those "powers which an agent can either exercise or not at a given moment, even holding all prior conditions at that moment fixed" (Steward, 2020., p. 345). At any given moment a song thrush, for example, can choose whether or not to exercise its power of vocalisation, either to sing, or to remain silent, *ceteris paribus*—the action is voluntary and so completely 'up to' the agent itself. Two-way powers are to be distinguished from one-way powers—those typically possessed by inanimate objects and substances—which are characterised by their possession of both necessary and sufficient conditions of exercise. Such powers cannot help but be exercised under the correct conditions. A golf ball cannot help but fall to the ground once in flight and thus it cannot be said that it is in possession of agency: matters are settled without any intervention on behalf of the golf ball itself. This is in contrast to two-way powers, where the agent controls both their initiation and their unfolding—actions are not agential purely because they are brought-about by the agent, but also because the agent retains the capacity to actively intervene and control the action as it is occurring. The agent is in control of their own actions "in the way that a government minister is in charge of a department" (ibid., p. 51)—if necessary, the agent can "step in" (ibid) and take direct conscious control over their own actions. From this, we can see how this conception of agency seems to resist universal determinism. If universal determinism were true, there would be nothing for an agent to *do*—agents would simply be "a mere part of the maelstrom of mere happenings, and [thus] would disappear from the world" (ibid., p. 155), and thus, because agents *do* exist—there is little doubt that there are self-moving animals in the world—it must be so that universal determinism is false.

There are two ways in which one may criticise Steward's Agency Incompatibilism. The first attacks to the metaphysical requirements needed for Agency Incompatibilism to operate and the second attacks the image of agency that it advances. We shall approach these criticisms in turn, starting with the flaws in the metaphysical position that Agency Incompatibilism requires to operate before moving on to the image of agency it presents, in an effort to show that the denial of Agency Incompatibilism is not only possible, but an attractive position to take.

3.2 PHYSICAL LAWS AND THE OPEN WORLD

If Agency Incompatibilism is to hold, there needs to be some kind of conceptual space in which the agent can dwell—a gap within the metaphysical schema that can be filled with an agent. Universal determinism holds that this gap does not, and cannot exist, without either falling foul of causal over-determination—being caused by two or more causally sufficient things simultaneously, or casting the agent in an epiphenomenal role with no real control of their own actions. For universal determinism, if the agent exists, they either have no impact on the causal chain, or they can only 'do' something that has already been done for them by physics. Universal determinism thus relies upon "the joint acceptance of the following four claims: (i) physical causal closure, (ii) causal exclusion, (iii) mind-body supervenience, and (iv) mental/physical property dualism (Kim, 2005., pp. 21 – 22). The synthesis of these claims suggests that an action—as it is material—must have only one cause (lest it be over-determined), and that cause can only be the material supervenience base of our mental states, and, as this base is material, it is settled by the laws of physics and not any intervention by the agent themselves. Free will therefore, cannot exist under this metaphysical schema, for the agent can have no proper impact on the unfolding of events, which are themselves settled mechanistically by material goings-on. In this materialist context, all an agent can be is an arena in which deterministic events unfold,

and thus it disappears into Steward's maelstrom of happenings. The future in the world of universal determinism can be thought of as being *closed*. No actions that are not mandated by physics can occur, so the future is settled and unitary—nothing can occur beyond that which is mandated by physical laws.

But as Steward argues, Agency Indeterminism requires that the future be *open*—"more than one future [must be] genuinely physically … possible" (op cit., 2012., p. 13)—but it is difficult to see how this may be the case if one material situation flows into the other along the mandates of the laws of physics. Steward's solution is to endorse radical indeterminism. She does not deny that the material states that our mentality supervenes upon must conform to the laws of physics, but she does deny that these laws are "*dictators* of reality" (ibid., p. 232)—she instead favours an understanding of physical laws as "*constrainers* of reality (ibid)". Physical laws therefore do not determine the precise path of the unfolding chain of events so that one link mandates the next, but rather physical laws act to determine what the next link along *may* look like: physical laws determine the repository of future potentialities available to the agent, but without mandating that one of these potentialities *must* be realised. The unitary, closed future of universal determinism is instead replaced by a plural future where "certain possibilities [are] left open by the world" (ibid., p. 126). In such a situation, the conceptual space for an agent to exist can be found without falling foul of the charges of epiphenomenalism and over-determinism. The agent is not an epiphenomenon, nor does it simply provide a second, over-determining cause because it can *do things that otherwise would not have been done*. Through the agent's own bodily movements, they can choose to realise one of the several available possible futures that do not violate the laws of physics. The laws of physics constrain available choices of the agent, but beyond that the agent is free to settle things how they wish.

From this, we can see that the existence of the agent, at least in any meaningful sense, is owed fundamentally to the understanding of physical laws that one may have. If one takes the view that physical laws are dictating then it is hard to see how anything like free will can exist, but if one takes the view that physical laws are constraining—as Steward does—then we can see how a free agent may be metaphysically possible. But why should we favour constraining physical laws over determining physical laws? Steward's argument for why we should take the latter view is simple. The former position is owed, in her view, to "extrapolating from the closed and ideal systems that physics and mechanics give us to think about the huge complexity of the real world—in ways that are inadmissible" (ibid., p. 231). Our successes in these ideal systems have prompted us to assume that we can apply the same rationale to the unideal and complex real world, so we come to think that we are bound by "a complete set of laws…that not only *constrains* but also *dictates* the entirety of what happens" (ibid). That such comprehensive laws exist is, to Steward, "sheer speculation" (ibid) and a product of "let[ting] our imaginations run away with us" (ibid), and so it would be wise to excise this assumption from our thinking. Such an avenue of attack though is misguided, however, for it is also true for Steward's own position.

The view that physical laws are constraining, rather than dictating is based upon the notion that the "laws that describe this world are a patchwork, not a pyramid … [with] pockets of great precision [and] large parcels of qualitative maxims resisting precise formulation" (Cartwright, 1999, cited in Steward, 2012., p. 231). Put simply, *dictating* laws assume that physical laws are pervasive and thus a complete and comprehensive understanding of *all* goings-on is possible with an appropriately complete and comprehensive understanding of physics, whereas *constraining* laws assume that there are some—most, in the opinion of Steward and Cartwright—areas of worldly goings-on that "occur by hap, subject to no law

at all" (ibid) and thus resist articulation via physical laws. Such a position requires that either a comprehensive and complete understanding of physics is impossible, or that free will always remains outside of a pocket of great precision and in the area of goings-on that resist articulation via physical laws (Beebee, 2014, p. 541ff). That this should always be the case is as much an assumption as that made by the universal determinist—it seems feasible that this situation may change as our investigations of the world progress. If Steward is able to dismiss the universal determinist on the grounds that their denial of free will is based on speculative premises, then we are just as able to dismiss Steward's position as it is built on similarly speculative grounds.

Both Steward's Agency Incompatibilism, and universal determinism are guilty of speaking with an authority that they are incapable of possessing: both incorporate an assumption about *future* physics into their respective premises. Each system is rooted in a *guess* of the ultimate nature of physics and thus, in the absence of a *God's eye view* whereby we could ascertain which party is correct, we have no reason at this stage to endorse either system over the other. If Agency Incompatibilism wishes to win our endorsement, it must do so on the grounds that its consequences are more attractive than its competitor, for as things stand, a choice between the two positions would be motivated by faith. Those with greater faith in physics would likely find universal determinism more appealing, and those with more faith in free will, are likely to find Agency Incompatibilism more compelling.

3.3 AGENCY INCOMPATIBILISM I: REACTIONARY FREEDOM

Prima facie, the metaphysical picture of Agency Incompatibilism is more attractive than its counterpart on the grounds that acceptance of Agency Incompatibilism requires no violation of our basic intuition that we are free agents. It seems obvious that, at each moment, we have a number of different avenues of action available to us—a number of

different futures from which we can choose—and the acceptance of Agency Incompatibilism leaves this intuition intact. Agency Incompatibilism is able to accommodate our intuitive possession of free will into its metaphysic (though it is perhaps more accurate to say that the metaphysic *requires* this intuition to be correct) and this is an obvious advantage over universal determinism, the acceptance of which forces us into denying that our intuitions are veridical. For the universal determinist, no matter how things may seem, there is no plurality of future possibilities—we are just the passive spectators over a mechanistically unfolding causal chain.

So it is so that Agency Incompatibilism is more attractive than universal determinism, but then we must also ask if Agency Incompatibilism is an attractive position to take, in and of itself, or is its attractiveness augmented by the unattractiveness of universal determinism? This is less clear if we focus on the image of free will that Agency Incompatibilism entails. Let us return to Steward's notion of settling. Steward maintains that agents are "constantly settling the answers to a variety of questions whose answers are … not *already* settled" (op cit., 2012., p. 39), but as we can see from this, there are *other* ways that matters can be settled too: "the actions of agents represent one means by which matters often come to be settled … [but] there may be types of occurrence other than actions which are able to be settled by certain matters" (ibid., p. 40).

Agency Indeterminism therefore posits the existence of at least two kinds of occurrence that can be counted as *settlers*—agential actions and non-agential events (ibid). It is with the interaction of these two kinds of settler that we can begin to see the unattractiveness of Agency Incompatibilism's understanding of free will. The distinction between action and event is motivated by whether the settling is owed to the agent or not—should the agent

intervene then actions settle matters, should no agent intervene then events settle matters instead. But if this is the case, it must surely be so that *events* settle matters through the operation of *dictating* physical laws, for if the laws were simply constraining then there would be multiple avenues of proceeding available and nothing to decide which would become realised—nothing could be resolved without agential intervention. Thus *events* must be vulnerable to determinism. The unfolding of happenings on a grand scale therefore must operate through the interplay between determined events and non-determined actions, but if this is so, and determined events are capable of making "hitherto open questions become no longer open" (ibid., p. 41), then it seems that at the point in time where *actions* settle matters, the agent must respond to the determined situation that has befallen them. The agent is thus un-free to perform any kind of settlings that are not mandated by the determined situation that immediately precedes their action. The free will of the agent therefore, does not extend to the point where they can *instigate* action purely by way of their own volition, rather, free will becomes necessarily *reactionary*—agency finds its articulation in *choosing how to react*. The agent must take up a role characterised by *response*, rather than by volitional *instigation* and thus the agency we find here is strange and unattractive.

3.4 AGENCY INCOMPATIBILISM II: PETTY FREEDOM

There are perhaps those who are not deterred from endorsing Agency Incompatibilism—reactionary freedom may be a price worth paying to be able to find a space for free will in a materialist metaphysic. Yet, the reactionary nature of Agency Incompatibilism's freedom is not its only unattractive feature. Its view of freedom is also marked by its *pettiness*. This is to say that the freedom that an agent can have, if Agency Incompatibilism is true, is not only reactionary, but fundamentally expressed in small actions, and although one can—as Steward does—argue that these small actions are a necessary part of larger actions,

situating freedom in these small actions leads to a claim that freedom exists in places where we would not intuitively say it belongs.

Consider Steward's example of a compulsive handwasher: an agent that, for one reason or another "simply *had* to wash [their] hands and could not leave it another moment" (ibid., p. 183). Intuitively, we might say that the agent in this case is not free in any recognisable sense of the term, but this is not what Agency Incompatibilism would have us believe. For Agency Incompatibilism, the compulsive handwasher is indeed free, even if they are unable to resist their compulsion to wash their hands. This is because the handwasher can still be thought of as making agential actions with regard to the precise *technique* of handwashing that they employ—the handwasher is still in possession of "freedoms to act in one way rather than another" (ibid). The handwasher can still settle a number of relevant matters; they "can wash with soap or not, with hot water or cold, for one minute or a bit longer, [etc.]" (ibid) and thus, for the Agency Incompatibilist, the handwasher is still free even when unable to resist their compulsions. This is an unusual conclusion, and one that would force us to find freedom in places where it would not intuitively be found. Sisyphus, for example, condemned to roll a boulder up a hill for eternity to atone for his deceitfulness is free to the Agency Incompatibilist because the technique he employs in doing so is up to him. Yet situating freedom in the atomic units of a complex action is problematic—there is a certain absurdity in claiming that the compulsive handwasher is free because they can roll their sleeves up, or that Sisyphus is free because he can choose precisely where to put his hands on the boulder—and so, as long as Agency Indeterminism is committed to the view that if *some constituent part* of an action is up to the agent performing it then they are expressing their free will, their position is tainted with this absurdity.

The image of freedom presented by Agency Incompatibilism is thus an unattractive one. The freedom of Agency incompatibilism is reduced to a choice of the *technique* by which we *respond* to the matters that befall us, and whilst this offers some slight improvement over universal determinism—which does not even afford us this petty, reactionary freedom—it does little justice to freedom as we may recognise it and so prevents us from endorsing Agency Incompatibilism on its own merits alone.

3.5 LIBET'S *A POSTERIORI* CASE FOR DETERMINISM

The failure of Steward's *a priori* case for free will is consistent with the conclusions drawn from scientific investigations into free will. Perhaps the most famous of these investigations comes from Benjamin Libet. Libet's experiments on conscious volition are straightforward: the subject of the experiment is strapped into EEG equipment that measures the electrical activity both in their brain, and in one of their arms and hands, and then, they are "asked to perform a simple quick flexion of the wrist or fingers at any time they felt the … desire to do so (Libet, 1985., p. 530). Such testing revealed that the electrical activity recorded in the brain consistently preceded the electrical activity in their arm or hand—an observation that can be explained by reference to the time it takes for electrical activity in the brain to be translated into electrical activity in the limb. Libet also notes that in each instance, the "subjects reported that they were aware of the urge … to act before every act" (ibid., p. 532). The subject was then asked to note when they first noticed their intention to move—the time of this was measured by recalling the position of a marker on a clock face at the time the subject became aware of their intention to move. These further experiments revealed that the reports of the subject's intention to move would occur *after* the electrical activity in their brain was recorded: the electrical activity that is responsible for moving our arm—the readiness potential—occurs about half a second before the movement itself, whereas the conscious experience of our intention to

flex our arm occurs roughly one-fifth of a second before the same movement (ibid., p. 529).

Libet's interpretation of his findings asserts that, because the readiness potential precedes the conscious experience of making the decision, voluntary acts are initiated unconsciously. When combined with Libet's belief that free will requires conscious volition, this prompts Libet to conclude that free will plays no part in our bodily actions (ibid., p. 536). Such a conclusion may be taken further to suggest that our conscious experience of purposely initiating the action is, in reality, "a post-hoc surplus process with no direct 'causal' efficacy in its own right" (Harnad, 1982., n.p.) thus meaning the phenomenology of the situation—i.e., that such an action is initiated by us—is illusory. Libet does not endorse this illusionist conclusion. Rather, Libet tries to rehabilitate the notion of conscious control by way of rethinking its role in bodily movement. The standard understanding of conscious control of our actions—the understanding suggested to us by our experience—is one where the role of conscious control is that of *initiation*. Libet's revised role for conscious control sees it not as the initiator of action, but as that which can *arrest* an action that has been initiated unconsciously—in the fifth of a second after we become aware of our intention to act, we can step in and abort the action, if we so choose: "the volitional process, initiated unconsciously, can either be consciously permitted to proceed … or be consciously 'vetoed'" (op cit., 1985., p. 537). Libet went as far as to test this theory—he ran a follow-up experiment where participants were instructed to *prepare* to flex their arms when the marker on the clock face reaches a particular position, but then to abort the intention to flex and remain still. The results of this experiment showed that electrical activity in the brain began to rise roughly a second before the marker reached the agreed position, only to drop about one-fifth of a second before the marker reached the

same position; Libet used these findings to evidence his suggestion that conscious control takes the form of a veto, rather than an initiation (ibid., pp. 536 – 538).

Libet's experiments seem to leave free will in a difficult place: they suggest that our conscious volition exists only as the overseer of unconsciously initiated processes and that free will as it is traditionally thought of does not match up with what is really going on in our brains. This though, is not necessarily the case. Though Libet's experiment is renowned and often relied upon to add useful *a posteriori* weight to determinist arguments, it is not without its flaws. It is possible to criticise Libet's conclusions from both a scientific and a phenomenal perspective. We shall do both, starting from the scientific perspective.

3.6 AGAINST LIBET I: THE SCIENTIFIC PERSPECTIVE

Let us recall the sequence of events that Libet's experiment suggests. If we say that time *t* is the movement of the arm, then at *t* -0.5 seconds, electrical activity in the brain begins to rise. Then, at *t* -0.2 seconds, the subject reports that they are aware of their decision to move their arm. Libet claims that the initiation of the subject's movement begins at the point where the electrical activity in the brain begins to rise (i.e., at *t* -0.5 seconds)—"neuronal activity associated with … the act has started well before any … conscious initiation" (ibid., p. 536)—or, to put it into different terms, the "brain 'decides' to initiate … the act before there is any reportable subjective awareness that such a decision has taken place" (ibid). Thus, the start of the increase of electrical activity in the brain at *t* -0.5 seconds is taken by Libet as the indication that the decision to move one's arm has been made.

Yet, there does not seem to be any reason to interpret the change of electrical activity in the brain at *t* -0.5 seconds as the decision itself. It could be that the initial rise of electrical activity is not the decision, but rather the start of a preparatory process that then leads to a decision being made at some point in the half a second between the initial rise in activity and the movement of the subject's arm. To dismiss such a possibility, an experiment would need to be undertaken to check if the initial rise in electrical activity can occur *without* the subsequent arm movement—yet Libet's first experiment did not do this; records of electrical activity were only made after the arm movement occurred (Mele, 2014., pp. 12 – 13). It is possible to argue that Libet's experiments into the conscious vetoing of brain initiated actions constitute such an experiment, and indeed, the results of this experiment seem to be useful here—a rise in electrical activity occurs until it declines roughly 200 milliseconds before the action would have been made—but ultimately, this experiment is not a suitable substitute for the one needed to dismiss the possibility that the decision to move one's arm occurs within the half a second between the rise in activity and the movement itself.

This is because Libet's experiment into the vetoing power of volition explicitly doesn't deal with *spontaneous* voluntary actions—the subjects are told when they should be (not) moving their arm. Moreover, this experiment is flawed as Libet could never guarantee that the subjects ever actually intended to move their arms. It does not seem possible for a subject to genuinely intend to move their arms in an experiment where they are explicitly told that such an action is to be aborted—there is an important difference between a situation where one genuinely intends to move their arm before then spontaneously deciding not to and a situation where one begins with the understanding that one's intention to move their arm should not translate into actual movement (ibid., pp. 17 – 19).

There are also doubts as to whether the results collected by Libet are sufficient to support his conclusion. Libet concludes with the strong claim that his findings show that "every conscious voluntary act is preceded by special unconscious cerebral processes" (op cit., 1985., p. 536). This generalisation seems to ignore an important distinction that can be made between different kinds of decisions; specifically, the distinction between deliberative decisions and spontaneous decisions. The distinction between a deliberative and a spontaneous decision can be understood by comparing the processes associated with making decisions that are important to those that are not—for instance, it does not seem that the quick flick of a wrist belongs to the same category of decision as buying a house or choosing a career. The former, a spontaneous decision, is characterised by a lack of reasoning—one simply flicks a wrist without giving it much thought—whereas the latter, a deliberative decision, is (usually) characterised by a comprehensive engagement with reason: one actively works out what is the best decision for them to take. As such, Libet's conclusion, due to the fact that the participants were explicitly told not to engage with any conscious reasoning, can at best only ever make the claim that *spontaneous* actions are unconsciously initiated: the assertion that *every* conscious act is unconsciously initiated is not supported by the results of Libet's experiments (op cit., 2014., pp. 13 – 16).

Finally, Libet's findings do not sit neatly with other similar experiments that investigate the relationship between movement and electrical activity in the brain (ibid., pp. 19 – 22). The most striking example of this comes from an experiment from Haggard and Magno that explicitly replicates much of the set-up of Libet's own experiments. The subjects are strapped into an EEG and a clock face with a moving marker is placed before them—some time after the subject presses a button to start the marker's travel around the clock, a high-pitched tone is played for a few hundred milliseconds. A short time after, this tone changes

into a louder lower pitched tone. This is the go-signal. The subject is tasked with pressing a button as quickly as possible after the go-signal is played before reporting verbally the location of the marker at the time they pressed the button (Haggard and Magno, 1999., p. 103). The results of this experiment revealed that the average time between the onset of the go-signal and the subject's movement is around a quarter of a second (ibid., p. 105). This suggests that Libet's assertion that the decision to move occurs roughly half a second before the action itself when the electrical activity in the brain begins to rise is incorrect. Instead, such findings seem to support the suggestion that the decision occurs at largely the same time as one becomes aware of their decision to act—Libet's experiment found that such awareness occurs roughly 200 milliseconds before the movement, within 50 milliseconds of the findings from Haggard and Magno. At best, this goes some way towards supporting a conclusion precisely opposite to that drawn by Libet—i.e., that the conscious volition *does* play an important role in decision initiation—and at worst, it demonstrates that the premises underpinning Libet's assertions are erroneous. In either case, the persuasiveness of Libet's arguments becomes diminished to such an extent that we can dismiss them; especially when taken in the context of the other objections thus far mentioned.

3.7 AGAINST LIBET II: THE PHENOMENOLOGICAL PERSPECTIVE

As mentioned prior, it is also possible to criticise Libet's experiments from a phenomenological standpoint. Libet's experiment involved participants voluntarily flicking their wrist, and measuring the discrepancies in time between brain activity, the conscious intention to act and the movement itself, which suggest that our brains decide to act before our conscious intention to act arises. In the experiment, the flick of the participant's wrist constitutes the action in question. Yet, it does not seem that, when proper attention is applied to the context of the experiment, a simple flick of the wrist is significant enough to

warrant the label of action, in and of itself. Rather, it seems that wrist "movement was itself only a minute part of a long sequence of movements amounting to a large-scale action" (Tallis, 2011., p. 248).

To illustrate this, focus your attention on the participant. There was a point in time, perhaps months prior to the moment that they flicked their wrist in Libet's lab, that they *decided* to participate in the experiment. Perhaps they responded to an email confirming their intention to participate, circled the date on their calendar and waited for the day to arrive. Then, when the day came, they left their house, took the train to the lab, sat in a waiting room browsing some magazines before then being called in to get wired into the EEG and start the experiment. The examination of these acts reveal that the flick of the wrist is simply the last stage of realising the participant's goal of contributing to Libet's experiments: the flick of the wrist is the final act of a decision initiated potentially months before the actual act took place (ibid). Libet's experiments, by virtue of their design, miss out the "temporal depth" (ibid., p. 249) of the choice to participate in them; they present a small movement—a simple component of a complex choice—as a substitute of a full-scale decision. This is of dubious utility: a part is rarely a valuable substitute to the whole which it belongs because the very act of choosing the part means that some potentially useful detail that pertains to the whole is neglected. In this case, it is the wider context that demonstrates that the decision to participate in the experiment (and by extension, the decision to flick one's wrist whilst strapped into an EEG) happens not only consciously but far before the event itself. The absurdity of taking a flick of the wrist as a representative of a full-scale decision can be shown by highlighting that fact that if simple movements count as decisions proper, then the walk to meet one's friends in the pub is not just one macro-level choice to walk to the pub, but is actually several hundred decisions—each step, each

obstacle avoided, each time a door-handle is used, etc., are the results of individual decisions (ibid). The lack of intuitiveness here is obvious, and because such a way of thinking underpins the conclusions that are drawn by Libet, we are again justified in dismissing them.

3.8 THE COMMON ERROR: A NEGLECTED FACTOR

With the image of free will presented by Steward's Agency Incompatibilism shown to be unattractive, and with the failure of Libet's *a posteriori* arguments for determinism, we find ourselves in a strange position. It is sensible to think that should either Steward or Libet fail in their arguments, then the other would be successful, but this is not the case—there are other positions that one can take in the debate. The failure of deterministic thinking (such as Libet's) creates sympathetic conditions for Agency Incompatibilism to flourish, but as we have seen, the best it can do with these conditions is to advance a conception of free will that is reactionary and petty. The dual failure of opposing viewpoints is a symptom of a more serious complication that underlies both arguments—a common error made by both parties that acts to render their view uncompelling. To uncover this error, we need to see what the two views have in common. There are two commonalities between the views that should interest us—the first is metaphysical and the second, relates to taking the position of detached observation to be the appropriate marker of truth in matters of free will.

Inspection of the two arguments reveals that Steward and Libet both operate under the auspices of a materialist metaphysic. Steward is upfront with this—the *raison d'etre* of her work is to carve out a space for free will within a materialist framework—and though Libet does not offer such candour about his metaphysical commitments in his writing, the fact that he is a scientist, in combination with his assumption that free will can be isolated

within neural happenings leaves his materialist leanings in little doubt. His experiments would make little sense if they were not grounded in the belief that material goings-on in the brain are an analogue of mentality. This shared commitment to materialism offers us an important clue in diagnosing the common error made by both parties. Recall the suggestion that IMF is the product of a distorting investigative methodology—a methodology that deigns truth to be always the product of an objective and detached scientific neutrality. Steward and Libet also employ the same kind of investigative methodology. Steward uses the detached observation of the self-propelled movements of non-sessile animals to ground her assertion that the world must be open, and thus free will has space to exist, and for Libet, the start of a measurable rise in electrical activity in the brain supposedly signifies that an unconscious decision has taken place. By accepting materialism and the supposedly neutral perspective of detached observation, both Libet and Steward are forced into understanding free will in material terms—specifically, in the material goings-on in the bodies of an agent. They both think that free will can be captured within bodily control and the physiological processes that underpin such; free will, for both parties, is to be found *in* the part of the material world encased in the agent's skin.

This does not *seem* to be too controversial. If we are to try and find free will, then it makes sense that we should look for it in free agents—it is absurd to look for free will in places in which it clearly cannot exist. But to locate free will *in the body of an agent* means that we misunderstand the actual nature of free will. Steward and Libet fail to understand "the proper level of description relevant to free will" (Gallagher, 2017., p. 145). This is obvious should one look at the everyday level of description that we ascribe to free volitional acts.

Say we were working in our gardens, digging a hole in the earth when we are stopped by another, asking what it is that we are doing. Our response would not be '*I am settling the matter of digging a hole in my garden through the exercise of two-way powers over my own bodily movements'* as Steward would perhaps suggest, nor would we follow Libet in saying *'I have permitted the neural activity initiated unconsciously in my brain to continue, and this has manifested itself in my digging of a hole'.* We would instead answer something along the lines of '*I'm planting a rose bush—I think it would be nice to see it from my office window'*. The relevant physiological goings-on inside our bodies are the furthest thing from our minds. Though this may appear pedantic—one rarely employs theoretical descriptions in everyday life—this response betrays the true level of free agential action—that "of a consciousness … *embedded* or *situated* in [a] particular context" (ibid). Both Steward and Libet *neglect* this embeddedness and only pay it attention insofar as the world is the arena in which an agent can act, but this is problematic as "the exercise of free will cannot be captured in or reduced to a description of neural activity or muscle activation or mere bodily movement" (ibid., p. 147). Our actions are free not only because they are in dialogue with the worlds in which we find ourselves, but also because they are expressions of our consciousness—the kinds of processes which Steward and Libet are concerned with are "carried along by … what is best described on a personal level" (ibid).

Steward and Libet do not do justice to this idea. Consider the wider context of Libet's experiment. A call for participants was likely circulated, and of all those it may have reached, we can assume only a few would respond. Indeed, it seems sensible to assert that the distinction between these two groups—those who respond and those who did not—is delineated by those who *made a choice to respond*, and those who *chose* not to. Moreover,

it is sensible to assert that this choice is best captured by a personal-level description—there is a *reason* that each group made their respective choice that cannot be captured in the physiological goings-on in their bodies. Perhaps the responders find the prospect of participating exciting, and maybe those who decline the invitation are busy looking after their children, or are working late in the office. One may look all one likes at neural occurrences, or bodily movements but one will never find these reasons because they only exist in the interplay between the consciousness of the agent and the world in which they dwell. The decision to participate or not, carries with it an "'existential' extensity" (Tallis, 2011., p. 249)—that is, to say that the choice is replete with "an explicit purpose, which pulls us towards goals we have ourselves envisaged and articulated" (ibid., p. 251). Yet to ignore the world in which the agent is embedded, or the consciousness of the agent, is to blind oneself to these key constituents of free action, which itself causes a misrepresentation of the complex and holistic nature of free choice. Free will instead is seen as discrete and atomistic, simply because the disinterested and neutral *langue* of scientific investigation is keen to regard subjectivity as superfluous to the truth.

We also find this disregard for the embedded, existential aspects of free action in Steward— such a disregard is the error common to both arguments—when we reconsider her master argument. Steward argues that the sheer existence of self-moving animals is enough to undermine universal determinism, and thus implies that the power to move voluntarily is the true seat of freedom, yet nowhere here do we find any attention paid to *why* the self-moving animal may *choose* to move as it does. We can recall that the compulsive handwasher is free because they may roll up their sleeves before succumbing to their compulsion, but Agency Incompatibilism is unable to grapple with the existential weight that such an action may carry. Indeed, this is not just an oversight on behalf of

Steward, but a *feature* of Agency Incompatibilism. We must remember that Steward maintains the position that human agency is only a problem for free will insofar as it belongs to the same kind of free will that is, she argues, also found elsewhere in the animal kingdom—it is "*animals* [that] make trouble for free will" (op cit., 2012., p. 3), not humans. Steward considers existential aspects of free will to be a dispensable relic of "an era which had yet fully to embrace the idea that human nature is continuous with that of the rest of the animal kingdom" (ibid., p. 2) and so begins her exposition of Agency Incompatibilism by explicitly moving away from a form of agency characterised by "rationality, deliberation, and forethought … creativity and even … self-development" (ibid), as this understanding of free will "already overestimates … the complexity of the phenomena required to make potential trouble for determinism" (ibid., p. 3).

By making this move Steward closes the space between humankind and our animal relatives and thus blinds herself to the fact that existential questions necessarily bear upon *free human actions*. This is the reason that we find her conception of free will unattractive. The view of free will Steward articulates, marred by reactivity and pettiness, is the kind of freedom we may expect an *animal* to have, and naturally, it is impossible to see how this kind of freedom is suitable for a *human* agent. We are not, as Steward seems to assume, just another kind of animal. We are unique in the animal kingdom precisely because existential questions bear upon our actions in a way that cannot be found in animal actions—a terrier does not dig a hole in the earth in order to plant a rose bush to gaze upon from its office window, it digs a hole in *response* to its predatory instincts.

From this we can see how the image of humankind presented by materialism is not just impoverished with regards to its neglect of consciousness, but also in terms of free will.

Under the auspices of a materialist metaphysic, there are two choices for explaining our free will: it is either *animal-like*, i.e., petty and reactionary, or it is illusory. No matter which the materialist chooses to assent to, the image of humanity is impoverished as a result—we lose our ability to recognise the existential significance of our actions and our choices—and as with consciousness before it, this again acts to limit the ways in which our technologies can be developed to be like us: if the blueprint can do no justice to our free will, then it will be lost in technologies developed in our image.

4 THE IMPOVERISHED IMAGE III: SEARCHING FOR SELFHOOD—BRAINS, STORIES AND REPRESENTATIONS

In the last section, we saw how materialism and its distorting methodology causes difficulties in accounting for human free will. Similar problems are also found when one tries to account for the existence of the self under a materialist metaphysic. That we *are* a self—alongside consciousness and free will—is one of the basic intuitions that we possess: we do not need to be convinced that we are selves because it is self-evident; it is through our own eyes that we see the world and it is *I* that cares about what happens to *me*. Yet, the materialist has difficulty in accounting for this—the objective world of matter and forces, seen from an Archimedean point has no obvious candidates for an entity that can be thought of as a self.

There are those, most notably Locke, who have tried to account for the existence and endurance of a self by appealing to our mental faculties, namely, our memories: "consciousness always accompanies thinking … [it] distinguishes himself from all other thinking things … and as far as this consciousness can be extended backwards to any past action or thought, so far reaches the identity of that *person*" (Locke, 1689., ii.xxvii.9). But, aside from being objectionable for a number of reasons—of which amnesia and paramnesia are perhaps the most striking (Flew, 1951., pp. 56 – 58)—appealing to our mental faculties *in and of themselves* is unacceptable to the materialist as it ignores the basic materialist tenet that all mental phenomena must either supervene on the material or be material in their own right. The uptake of this tenet has prompted materialists to look at the brain to find a material basis for selfhood. There are two types of materialist theories of self that arise from this—those that say that the self is numerically identical to some material feature of the world, and those that say that there is no material feature of the

world that could be a self, and so, the self does not actually exist and is thus illusory. We shall begin by exploring an exemplar of the first type—Peter Van Inwagen's assertion that the self is numerically identical to the brain—before turning our attention to two illusionist views of selfhood, one from Daniel Dennett, who claims that the self is a narrative construct, and another from Thomas Metzinger, who claims that the illusion of self arises as a consequence of representationalism.

4.1 ME, MYSELF AND MY BRAIN

It is intuitive to assume that psychological continuity is an important factor that a reasonable account of selfhood must account for—it is difficult to see how a person could be the same person that they were at the beginning of the week if by the end of it they have begun again with a *tabula rasa*. But, as we have mentioned, appealing to psychological continuity alone does not satisfy the materialist as it pays no attention to the material basis of this psychological continuity. Such observations have led to the suggestion that the existence and the endurance of a self over time is contingent upon the existence and endurance of the brain, insofar as it is responsible for the relevant kinds of phenomena associated with psychological continuity—memories, opinions, and the like.

Peter Van Inwagen is a proponent of such a position. He argues for this position first by asking us to "suppose one of [our] fingers has been cut off" (Van Inwagen, 1990., p. 169). Such a wound, Van Inwagen rightly suggests, would have no impact upon our self-hood—we are as much of a self with nine digits as we are with ten.

This is also true if we are to suppose we lose an arm or a leg—"a person who suddenly loses his limbs … is maimed but continues to exist" (ibid., p. 170). Van Inwagen continues, asking us to consider the selfhood of a severed head "arranged [in] an elaborate mechanism to keep [it] alive" (ibid). Again, Van Inwagen suggests that the severed head

would still be a self; he even goes as far to suggest that a "naked brain" (ibid., p. 172), or "a *part* of a brain" (ibid) linked into a life support system would be enough to maintain the selfhood of the person who's brain it is. We can pare away the entire body—or the "brain-complement" (ibid., p. 173) in his terminology—without ever interrupting the selfhood of the person in question. It is "a trivial truth" (ibid., pp. 175 – 176) in Van Inwagen's view, that where there is a brain, there is a self. Though not a proponent of the position himself, Derek Parfit notes that outsourcing our selfhood to the brain as Van Inwagen does, has some intuitive attractiveness. If we grant, for argument's sake, that isolating and excising the brain (or the relevant parts of the brain) is possible, and they may be incorporated into some kind of communications system, we may believe that a self has endured should it be able to demonstrate that psychological continuity remains—if both the donor of the brain/brain-part, and the brain/brain-part-plus-communications-system could signal that *Ward No. 6* is their favourite Chekov story we may draw the conclusion that the same self remains (Parfit., 2012., p. 17). And similarly, it also offers a plausible explanation of why we may see brain-death as the death of the self (ibid., pp. 17 – 18).

Locating the self in the brain, or in a brain-part though, is problematic. Firstly, paring off a part of human physiology in an effort to find the smallest unit of the body that could be called a self, as Van Inwagen does, mistakes the nature of the human being. The body is not a collection of discrete parts to which tasks and functions can be delegated in such a way that the function is maintained in isolation away from the whole. It is a holistic *organism* of interconnected and symbiotic parts that rely on each other for their utility—even if it was possible to isolate a self-laden brain-part "its function cannot be separated from that of the remainder of the brain plus all the bodily structures that feed into and out of the brain" (Tallis, 2020., p. 159). Van Inwagen, therefore, cannot point to a discrete

section of our physiology and say that it is the part responsible for the self, because even if it was so that we could reduce mental phenomena to the brain (or a brain part), that specific part of our physiology could not support a self *outside* of the bodily organism—it needs the rest of the human organism to support its functions. Though Van Inwagen may object, other similarly inclined philosophers could argue that such an objection is taking the point too literally, and what is actually meant is that selfhood can only be accounted for in a *total* organism in possession of the correct brain-part. Such a position would share considerable conceptual ground with Eric Olson's animalism, the view that "each of us is numerically identical with an animal: there is a certain organism, and you and it are one and the same" (Olson, 2007., p. 24), but this view is equally misguided albeit in a different way, as it reduces selfhood simply to physiology and pays scant attention to the role that our mental lives play in matters of selfhood.[16]

Secondly, it does not seem that a disembodied brain or brain-part "would have the memories and thoughts of a living person" (Tallis., 2020., p. 159). For example, the disembodied brain/brain-part—unable to receive or parse any sensory information—could not, for instance, look at the autumn leaves and think how they remind them of the red of a post-box. Nor does it seem that a disembodied brain/brain-part could remember how to perform practical tasks that resist translation into maxims—should the brain have belonged to a pianist, it is hard to see how they could remember precisely how to play their favourite piece of music.

[16] As Tallis highlights, identifying the self with our bodies also entails that any mental actions, even those associated with purely abstract entities, e.g., maths, would be carried out by our entire bodies—a prospect that he suggests "will be flattering our kidneys" (Tallis, 2020., p. 159).

Thirdly, identifying a self with a piece of matter means that the self would also lose access to its temporality. If the self can be pared off from the body, as Van Inwagen suggests is logically possible, then that self could exist without any further input: there is no way to get anything—sensory information, propositional knowledge, etc.—to the disembodied self. This means that the self would be forced to reckon only with the experiences that the self has had *prior* to being pared away from the body, effectively voiding the self's ability to reckon not only with its present, but also with its *future*. The isolated brain/brain-part self could perhaps exist—with the aid of thus far unrealised, but feasibly possible life-support systems—but its phenomenal perspective would be necessarily retrospective in nature. This sits in contrast to selfhood as we experience it: alongside reckoning with the past, we can also reckon with the present and the future—we can deal with what is before us, and we can anticipate what will come afterwards. As Van Inwagen's brain-self suggests that a self can exist in a purely retrospective mode it does not seem that it does justice to selfhood as we know it—in light of this, and the objections above, we are safe in dismissing Van Inwagen's theory that the self is numerically identical with either the brain or the supposedly relevant parts of the brain.

4.2 STORIES ALL THE WAY DOWN: DENNETT AND SELFHOOD

It is reasonable to assume that all theories of the self that argue that we are numerically identical with the brain, or one of its material features, in the way that Van Inwagen suggests, are vulnerable to the objections made above—should one peer into the brain, there is nothing in its material goings on that looks to be a suitable candidate for a self. This has prompted the idea that the self is illusory, non-existent in any material sense, and something that we are mistaken about when we say it exists. The materialist's task thus shifts away from finding an isolatable material self to explaining *why* we think that selves exist in a way that is sympathetic to their metaphysical commitments.

To explain the illusion, Dennett seeks to draw a parallel between the self and what he sees to be another "theorist's fiction" (Dennett, 1992., n.p.)—an object's centre of gravity. A centre of gravity is a useful concept. It can help us make predictions about and explain what happens in the world, it is something that we can manipulate by moving objects, and it is something to which we can ascribe a history. Yet, despite all of this, a centre of gravity does not really exist: "a centre of gravity is *just* an abstractum" (ibid).[17] Therefore, trying to identify a feature of the physical world that is numerically identical to a centre of gravity is "a category mistake" (ibid). The same applies also, in Dennett's view, to the self—the self is also a fiction, a story we tell about ourselves and others not only for the sake of its explanatory efficacies in accounting for human behaviours, but also as a way of offering unity and coherence to the discrete and atomistic events that befall us. For Dennett, these events are tied together through the creation of a coherent narrative. Our selfhood exists only as the central figure of this narrative—"we try to make all of our material cohere into a single good story … the chief fictional character at the centre of that autobiography is one's *self*" (ibid). If this is correct, then trying to find the self materially instantiated somewhere in the world is making the same kind of category error as one who seeks to find the material instantiation of a centre of gravity. The self does not exist beyond the *I* that features in our own autobiographical account of the world—it is an illusion that we employ to confer coherence to what happens to us.

There are two problems with Dennett's illusionist view of the self. The first is that for Dennett's argument to make any sense, the existence of a self must first be presupposed. We can see that there is something unusual about the analogy between the self as a

[17] Dennett makes a similar point with his idea of a "lost sock center" (Dennett, 1991b., p. 28); another abstractum "defined as the center of the smallest sphere that can be inscribed around all the socks I have ever lost in my life" (ibid). In his view, both a centre of gravity and the midpoint of the inscribed lost sock sphere have the same metaphysical status—a non-existent abstraction.

theorist's fiction and a centre of gravity as a theorist's fiction—should one pay attention to the 'theorist' in each case, this is clear. A centre of gravity is dissimilar to the self, insofar as the theorist in questions concerning an object's centre of gravity is external to the object itself, whereas the theorist in question of matters of selfhood *is exactly that self*. A centre of gravity is not something that an object "ascribes to itself but something that we … ascribe to it" (op cit., 2020., p. 150). An object does not undergo any self-interpretation to arrive at the position that it is in possession of a centre of gravity—rather, that is something that *selves* do to *it*. The existence of the self is in the background of Dennett's theory as that which interprets worldly entities, itself included, Moreover, by even suggesting that the self may be illusory, Dennett is demonstrating that selves exist (precisely, that *he* is a self), for "if the self is a fiction, it is one which I—presumably a non-fictional I—propose and embrace" (ibid). Selves simply would not show up as an issue in a self-less world.

The second problem is that the image of a fictional self does not do justice to selfhood as it—unsurprisingly, considering Dennett's hostility to the notion—leaves out the phenomenal *experiences* of a self. We can concede Dennett's point that events can happen to a fictional self, and certain facts can hold about a fictional self, but it does not seem that a fictional self can *experience* anything: we can say that Raskolnikov is a murderer, but it is another to say that he actually felt guilt afterwards—there is no real guilt, because there is no real Raskolnikov. This stands opposed to our own basic conception of our selfhood—it seems obvious that we *do* experience things, and that these experiences are intrinsically linked to our selfhood: "an experience is impossible without an experient" as Frege noted (Frege, 1918., p. 299) and since a fictional self cannot experience anything, the fact that we *do* suggests that we are non-fictional selves.

From these criticisms we see that Dennett, at best, fails to offer an account of selfhood worthy of the name, i.e., one that is capable of accounting for phenomenal experience, and at worst, actually demonstrates the existence of a non-fictional self—thus Dennett's argument, like Van Inwagen's, can be safely dismissed.

4.3 VIRTUAL REALITY ME: METZINGER ON SELFHOOD

Claiming that the self is numerically identical to the brain or one of its parts, or that the self is, in essence, nothing more than a narrative construction are not the only options available to materialism when it comes to offering an account of the self. The materialist can also claim that our selfhood emerges as a consequence of the correct kind of process occurring in the brain. Such a position splits the difference between the two theories we have thus far explored. Like Van Inwagen, it suggests that the self is intrinsically linked to the brain (or at least, something the brain *does*)—so it satisfies the materialist criterion that all must either be material, or supervene on the material—but like Dennett, it also denies that there is a material entity in the world which is numerically identical to a self. If our sense of self arises as a consequence of a brain process, there is nothing that *is* the self, so again, the self does not actually exist. This means that it also side steps the kinds of problems that plague Van Inwagen's position. Let us look at this position with more detail.

Thomas Metzinger is perhaps the most prominent proponent of the claim that the illusion of the self arises via brain processes. He begins his argument by asking us to consider the rubber-hand illusion, revealed in an experiment by Botvinick and Cohen. The set-up of the experiment is simple: a subject is "seated with their left arm resting upon a small table [with a screen] position beside the arm to hide it from … view and a life-sized rubber model of a left hand and arm was placed on the table" (Botvinick and Cohen, 1998., p. 756). The subject is then instructed to sit with "eyes fixed on the artificial hand whilst …

two small paintbrushes … stroke the rubber hand and the subject's hidden hand [simultaneously]" (ibid). After the experiment, the subject reported that they felt the sensation from the paintbrushes as located not on their real hand, but rather the artificial limb before them: "the illusion's spurious reconciliation of visual and tactile inputs relies upon a distortion of position sense" (ibid). It is supposedly this reconciliation between visual and tactile input that leads to the subject "'thinking it [the artificial limb] was actually [their] own'" (ibid). To explain this experience of unity between oneself and the artificial limb, Metzinger posits the existence of a "*phenomenal self-model"* (Metzinger, 2009., p. 4). The phenomenal self-model (PSM herein) is the "conscious model of the organism as a whole that is activated by the brain" (ibid)—when the rubber hand illusion holds, it is because, in Metzinger's view, the artificial arm has been incorporated into the subject's PSM: "whatever is part of your PSM … is endowed with a sense of 'mineness', a conscious sense of ownership" (ibid., p. 5). But positing the existence of a PSM prompts two questions—what exactly *is* a PSM, and what does it have to do with an account of selfhood?

To understand what a PSM is, we must look further into Metzinger's theory of consciousness. For Metzinger, consciousness is not only the appearance of a world, in a phenomenal sense—"if you are conscious, a world appears to you" (ibid., p. 15)—but also it is that which *situates* us in a world replete with various phenomenal experiences (ibid., p. 19). As Metzinger states, "when you wake up in the morning, you experience yourself as existing at a specific time, at a single location, and embedded in a scene" (ibid). This phenomenon of being *situated* in an experientially rich world is, in Metzinger's view, owed to the fact that the external world is represented internally on a neurobiological level: "conscious experiences are full blown mental models in the representational space opened

up by the gigantic neural network in our heads" (ibid., p. 23). Our sensory organs provide our brains with the raw information—"ultimately, our subjective experience is a biological data format" (ibid., p. 8)—necessary for the creation of a detailed mental representation of the external world; a representation that also includes a representation of our bodies. For Metzinger, "our brains generate a world-simulation … then, they generate an inner image of ourselves as a whole" (ibid., p. 7). It is this representation of our bodies, that exists within the representation of the world, that *is* the PSM.

Here we can begin to look at our second question—what does the PSM have to do with selfhood? For Metzinger, the PSM is not simply a representation of our organism, but it also possesses *content*: "your bodily sensations, your emotional state, your perceptions, memories, acts of will, thoughts" (ibid., p. 8) or, more straightforwardly, "the content of the PSM is the Ego" (ibid., p. 5). In addition, the PSM is not just a container for the ego, it is also a mechanism through which an "organism [can] interact with its internal world as well as with the external environment" (ibid., p. 4). It is a tool that, once generated by the brain, unlocks the organism's ability to interact with the world. By generating a representational model of the environment and a PSM, the brain "places you at the center of a behavioural space … your consciously experienced world-model" (ibid., p. 105) and thus allows an organism to "appropriate its own hardware" (ibid., p. 105) and *act* in their environment.

A valuable analogy is offered by Metzinger that helps us to grasp this point. Imagine that one is playing a particularly immersive and realistic video game. The world in the video game is not real—rather it is a *virtual* representation of "a *possible* reality" (ibid., p. 106) much alike to the internal representation of the human organism's environment generated

by the brain. It just so happens that one representation is realised by computer hardware, and the other by neurobiological means. Similarly, the character that one controls in the video game is also an unreal virtual representation of a (human) organism, much like the PSM generated by the brain—and, again, in much the same way that the in-game representation of a (human) organism opens up the in-game representation of the world as a behavioural space, where the in-game character can *do* things, the PSM allows the human organism to *do* things in their own internally represented environmental model. It is this notion of the PSM as a virtual centre within a represented behavioural space, replete with 'ego' contents, that gives rise to the illusion of the self. The PSM and the internal representation of an environment accounts for what Metzinger suggests is "minimal self-consciousness" (ibid., p. 102) i.e., that "what makes control *possible* … [namely] an image of the body … plus a spatial frame of reference … [and] a visual … perspective originating within the body" (ibid). As such the organism deceives itself into thinking the self is real, not only because it *can* act, but because its experiences of the represented environment are experienced from a *first-person perspective: "*you see *with* [the ego]" (ibid., p. 8).

4.4 PROBLEMS WITH METZINGER'S VIRTUAL SELF

There are a number of ways that one may criticise Metzinger's argument—we shall look at three such ways. Firstly, we shall explore Metzinger's arguments as to why our intuitive position on selfhood—i.e., that we are real, actually existing selves—is mistaken, and show them to be ineffectual. Secondly, we shall pay greater attention to his claim that *action* is reliant on representation, arguing that he is mistaken in this view. These two arguments seek to undermine the motivation for assenting to Metzinger's view by showing that its illusionist conclusions are based upon weak premises, and by showing that his representationalist position is unnecessary; ultimately, Metzinger provides no good reason to move away from our intuitions. Finally, it shall be shown that, like Dennett's

illusionism, Metzinger's view actually *presupposes* the existence of a real self, and thus its conclusion that selves are not real is incoherent.

4.5 AGAINST METZINGER I: THE JUSTIFICATIONS FOR TRANSPARENCY

Metzinger suggests that, by default, we are naïve realists. We think both that our selves have *actual* existence and that we are in direct contact with our world. In Metzinger's view, this is because our internal representations of our organism and our environment are phenomenologically transparent. We fail to realise that our existence is that of a high-level representational centre embedded inside a low-level representational behavioural space because the phenomena of our situation tells us otherwise: "even if we believe that something is just an internal construct, we can experience it only as *given* ... [because] we have no point of reference 'outside' [of our experience]" (ibid., p. 44). The real existence of the self, and the world *as it appears to us*, are nothing more than brain-generated illusions; we are only ever in contact with our representations. Metzinger offers two reasons to explain *why* we fail to recognise that our reality and our selfhood are illusory.

The first of these is the high speed at which the brain is capable of reliably processing the information sent to it by our sensory organs and updating the PSM and the behavioural space it is embedded in: "[the] global model is a *real-time* model; it is being updated at such a great speed and with such reliability that in general we are not able to experience it *as* a model anymore" (Metzinger. 2004., p. 555). Such a notion seems to be intuitively sound—if there is no breakdown in the representation, then there is no reason to see it as such, in much the same way that our pens do not show-up as objects until there is some kind of breakdown caused by their malfunction. So long as the brain is quick and reliable, then it is able to update its representations in such a way as to never bring the fact that they *are* representations to our attention. There are two problems with this position.

Firstly, it does not seem that our access to the world is all that reliable. Rather, it is more accurate to suggest that our experiences are usually imbued with ambiguity. There are many examples of this: anyone who has mistakenly heard their name uttered in a bustling café, or have waved to a stranger believing it to be their friend can attest that our grasp on the world is not as reliable as Metzinger may have us believe. Secondly, the notion of speed is not suitable for explaining the illusion. Say an individual has a neurodegenerative condition that causes their neural conduction velocity to drop over time, so the speed at which the brain can update the representation decreases over time also. If we follow Metzinger's suggestion that the speed at which the representation is updated is responsible for the illusion, then it seems as though there will be a point at which the illusion ceases to hold thus making the individual aware of the representational nature of their experiences. Metzinger would respond by arguing that there is no *objective* threshold value which if dipped below, causes the illusion to cut-out, as the illusion instead hinges upon "the speed of different types of processing … relative to each other" (op cit., 2009., p. 42). But this is little improvement—the threshold would still exist, but the value of that threshold would be *relative*, rather than objective. It is still logically possible for an individual to dip below this relative threshold and become aware of the representational nature of their experience. If one appeals to speed, in and of itself, in order to justify the illusion, then they must also offer some insight into the threshold value of speed above which the illusion can hold—if no such threshold is stipulated then *any* neural conduction velocity must be able to maintain the illusion, whether very low or very high, and thus they stop talking about *speed* and start talking about the brute presence of brain-activity instead. But if the

threshold *is* stipulated, then the argument is vulnerable to the notion that the illusion can cut-out if this threshold is crossed, which is obviously absurd.[18]

The second argument that Metzinger employs to explain the transparency of our representations appeals to human evolutionary history, specifically the notion of "*metabolic price*" (ibid., p. 43). Metabolic price is straightforward—if the human organism develops a new capacity, then the new capacity must be paid for with additional energy gleaned from the environment: "if an animal is to evolve, say, colour vision, this new trait must pay by making new sources of food and sugar available to it" (ibid). Metzinger argues that the recognition of our experience as representational would not have helped our ancestors offset the cost of such metarepresentations—knowing our experience is representational does nothing to help us find more food in our environment. The metabolic cost of recognising the illusion would remain outstanding, so we did not evolve the capacity to form these metarepresentations and thus our representations remain transparent. Or, furnished in different terms, "the formation of metarepresentations would not have been cost-efficient: it would have been too expensive in terms of the additional sugar we would have had to find in our environment" (ibid., p. 44).

Appealing to cost-efficiency is an unusual move for Metzinger to make. Recall that Metzinger argues that our selfhood and our conscious experience are the product of a robust and detailed internal model of our environments, realised on a neurobiological

[18] There is evidence in the scientific literature that neural conduction velocity is influenced by a number of factors, including age, height and bodily temperature (for example, Stetson, et al., 1992., pp. 1100 – 1102). For example, the evidence regarding temperature claims that lower temperatures are associated with lower neural conduction velocities, which, when taken in tandem with Metzinger's argument that speed plays a vital part in maintaining the illusion of naïve realism, suggests the logical possibility that a suitably cold individual may be able to recognise first-hand that their experiences are representational and they are nothing more than the content of one of these representations.

level—the brain actively constructs what we believe to be our reality. Moreover, it does not do so by capturing the world in a one-to-one simulation—Metzinger also argues that the brain *embellishes* the real world by imbuing it with phenomenal qualities that he suggests do not exist in the real world: "the apricot-pink of the setting sun is not a property of the evening sky; it is a property of the internal *model* of the evening sky, a model created by your brain" (ibid., p. 20). It is hard to see how the creation of a constantly-updated model of the world, embellished with qualia that have no real existence, could come at a lower metabolic price than the naïve realism that Metzinger argues against. A re-creation of an environment—indeed, a recreation that is *more* detailed than its original—is less cost-efficient than direct access to that same environment. Making a redundant and unnecessary neural copy of the information contained within the world comes at a higher cost than reckoning with the world first-hand.

Let us use a couple of examples to demonstrate how direct access to the world may come at a lower metabolic price than representationalism. Firstly, say one is playing a game of Tetris. There are two ways in which we can do this. One option, which is likely Metzinger's preference, is that we recreate the image of the screen and its contents in our brains and then, refraining from practice, we mentally orientate the falling piece to align it with the spaces below. We then act in such a way to make the on-screen piece align with our mental image. Another option is that we press a button to rotate the on-screen piece and assess its suitability in this practical manner instead.[19] As the second example requires no redundant mental representations of the game, it is the more cost-efficient option of the

[19] This example is loosely paraphrased from *The Extended Mind* (Clark and Chalmers, 1998. In Clark, 2008, p. 220ff.).

two—the information is already out there in the world, so reproducing it mentally comes at unnecessary cost.

Secondly, say one is at one's desk trying to solve some kind of difficult, cognitively-demanding problem by making notes on a piece of paper (Clark, 2008., p. xxvff). There are two ways in which we can look at this. One option—again, likely the one favoured by Metzinger—is that the working-out on the paper is little more than an external record of the mental work being done internally. Another option, is that the problem-solver's notes actually constitute the agent's thinking—the problem-solver's mental labour is done *through* their notes. Any jottings on the paper do not have to be actively held in their mind—they can be stored *out-there* and be recalled via sight. In this example, like the other, there is no need for the problem-solver to create any mental representations. All the information is stored *in the world*; bringing it *inside* is effectively making an unnecessary duplicate so would come with a greater metabolic cost than directly accessing the information already stored in the world. Therefore, Metzinger's assertion that metabolic price explains the illusion seems to be a *post hoc* justification for his existing representationalist proclivities rather than one of the principles on which it was founded. If it was so that the notion of metabolic price was a guiding principle in his position, Metzinger would likely not endorse representationalism, insofar as its metabolic price is intuitively higher than its naïve realist alternative.

From this discussion, we can see that Metzinger's rationale for moving away from naïve realism is flawed. Arguing that we are not aware that our world and our selves are simply representations because our neurones are quick, or because it would be too costly in terms of energy to see beyond the illusion is flawed. These arguments would only be persuasive

if we were—like Metzinger seems to be—already predisposed towards believing in representationalism, *viz* the "widely shared … scientific worldview" (ibid., p. 126) to which Metzinger surmises his representationalism to belong. As such, our assent would not be paid for in rationality, but rather in faith, so we are justified in asserting that Metzinger offers no good explanation as to why our intuitive assumption that we are in direct contact with the world is incorrect.

4.6 AGAINST METZINGER II: REPRESENTATIONS, ACTION AND OTHERWISE

Even if Metzinger was to concede that his justifications of the transparency of our representations are weak, he would likely not move away from his representationalist position on the grounds that he believes that our representations are necessary for action. He would argue that without representations, we would not be able to *do* anything because our environment offering itself up as a behavioural space is contingent on our internal representations of it and our bodies. Thus the very fact that we can act gives credence to his representationalism.

To criticise Metzinger's position, more needs to be said about his theory of action. It is easiest to understand Metzinger's theory of action if we think of actions as the product of an action generating process—a process which has a precise order of operation that can be broken down into three stages. The first of these stages we are familiar with.

A mental model of the human organism and their immediate environment, realised in the brain on a neurobiological level, is generated. Metzinger claims that this mental model is *online*, i.e., it is one that is "permanently being modulated by the information flow of the

sensory organs" (op cit., 2004., p. 51).[20] The generation of this *online* model allows the second stage to be realised—the *offline* simulation of the desired action. This simulation is undergone to isolate the anticipated *motor qualia*—the "simple forms of sensory content that are available for selective motor control, but not for attentional or cognitive processing" (ibid., p. 76)—associated with the desired action. As Metzinger claims: "we need an internal image of our body that predicts the likely consequences of, say, an attempt to move our left arm in a certain way … [and] to be really efficient, we need to know in advance what this would feel like" (op cit., 2009., p. 111). Once we have "internally simulat[ed] the expected *external* sensory consequences of a specific action" (op cit., 2004., p. 183) and isolated the anticipated motor qualia, we move to the final stage. In this stage, the body moves and is monitored by a "full-blown sensorimotor loop" (ibid) until the content of the *actual* motor qualia are in line with the content of the *anticipated* motor qualia, at which point, the action is completed. To put this process briefly and simply, before acting, our brains run a simulation of that act using the representational models that it creates, and it then moves our actual body in line with this internal simulation.

The problem with the above account of action, is that it presupposes the existence of entities—namely, internal neurobiologically-realised representations—that are unnecessary for a theory of action. If one returns to the phenomenology of action, then a serviceable account of action that carries a smaller ontological burden can be given. And moreover, if we can explain action without resorting to representation, then there is no reason to

[20] In short, a representational model is online if it is actively tracking the environment or the human organism, *as it currently is*. This is to be contrasted with an *offline* model, where the model does *not* represent the environment as it currently is. For example, Metzinger suggests dreaming is a "complex *offline* hallucination" (ibid)—the dream-model does not track the environment, or the human organism, as they currently are; if dreams were online, you would dream that you were asleep in bed, dreaming.

endorse Metzinger's representationalism and, *a fortiori*, Metzinger's argument that the self is nothing but the representational content of our PSM.

4.7 ACTION WITHOUT REPRESENTATION

The concept of *affordances* is central to non-representational theories of action. Affordances, put simply, are "what [the environment] *offers* the animal, what it *provides* or *furnishes*, either for good or ill" (Gibson, 1979., p. 119). These environmental offerings cannot be captured in material terms: they do not exist objectively in the world, nor do they exist in some representational internal space—rather they exist in the *relation* between the agent and their environment (ibid., p. 120). As such, the affordances present in one agent's environment may be different to those afforded to a different agent in the same environment. For instance, a parent and their toddler may be in the same environment, yet the child's chair, being made to fit their size, only affords the opportunity to sit to the child and not their parent. Under the theory of affordances, the parent does not recognise the unsuitability of the child's chair because they first run an internal *what would happen if I sat there* simulation beforehand, but instead, because the chair does not *look* like it is appropriate for sitting upon: "the affordance [or lack thereof] is perceived *visually*" (ibid., p. 120; emphasis mine). No representations are necessary because the relevant information is already *out-there* and thus requires no internal duplication. Instead the agent, embedded in their environment, acts by way of visually perceiving the potentialities of action available in that environment, before then choosing one of these and committing to it.

But what of the need to 'know beforehand' that Metzinger suggests is key to performing a successful action? There is some truth in Metzinger's position—being able to anticipate what will happen if a particular course of action is undergone is important to its successful execution—but where Metzinger falters is in his account of where this knowledge

originates from. There is no need to posit an offline representational test space for the purpose of performing *dry-run* simulations of actions—the knowing-beforehand need not originate from rationalising about one's actions. The skilled performance of actions can also (indeed, more commonly) originate from practice.

Though most of the skills that we frequently employ have been acquired through trial and error, or the observing and mimicking of others—for example, learning to walk and learning to talk—let us follow Dreyfus (2002., pp. 368 – 372) in giving a clear example of skill acquisition, to lay bare the intricacies of such. Let us borrow his example of learning to drive. The first step for the novice driver is to understand the theory of driving, and this is done through "decomposing the task environment into context-free features which the beginner can recognise without previous experience in the task domain" (ibid., p. 368). The novice driver is then given rules to follow that bear upon these context-free features—e.g., *shift gear when your tachometer reads 2000rpm*. After following such rules for a while, the novice driver will then begin to perceive other, context-dependent "aspects of the situation" (ibid., p. 369) that then can be used in place of the rule-based strategy that they used previously. Instead of consulting the tachometer, the novice driver can instead rely on the sounds of the engine to let them know when the engine affords to them an opportunity to shift up a gear. At the next level of familiarity with the task, the number of additional situational aspects increases, and thus presents the novice driver with the task of discriminating which of these additional aspects are appropriate for the task at hand—the novice must "restrict themselves to only a few of the vast number of possibly relevant features and aspects" (ibid). This is done, initially, through the application of further rules—(e.g*., drop to 30 when taking this turn because the ground is wet*, etc), generated on-the-fly by the novice themselves. By doing this, the learner begins to develop a *feel* for

the task—the "theory of the skill, as represented by rules and principles, will thus gradually be replaced by situational discriminations accompanied by associated responses … the learner simply sees what needs to be achieved" (ibid., pp. 370 – 371). Finally, once the learner has a well-developed feel for the task, he "not only sees what needs to be achieved … [but also] sees how to achieve his goal … [and thus] allows the immediate intuitive situation response that is characteristic of expertise" (ibid., pp. 371 – 372). They can operate intuitively through perceiving the opportunities for action afforded to them by the car and their environments.

From this example, we can see that two things happen. The first of these happens early on in the example—when the novice gets behind the wheel of the car. As soon as they experience what it is like to drive, they begin the process of incorporating the vehicle into their *intentional arc*. The agent's intentional arc is that "which projects round about us our past, our future, our human setting, our physical, ideological and moral situation" (Merleau-Ponty, 1945., p. 136) and thus "brings about the unity of the senses, of intelligence, of sensibility and motility" (ibid). Through incorporating the car into their intentional arc the agent becomes *situated* in the car—they begin to become involved in the business of driving by forging a tight link between their body, the car and the environment at large. They become *coupled* with the car by way of "incorporate[ing it] into the bulk of [their] own body" (ibid., p. 143). This coupling then gives rise to the second thing we can see happening in the example. Once coupled with the body, the car offers up a variety of situational aspects that must be reckoned with—the agent must begin interpreting which of these aspects are meaningful and thus demand our attention and those which are not: they must get a good *grip* on their situation. This is done by bringing certain features out from the background milieu of situational aspects into the attentional foreground, thus bringing

what is important in the situation into proximity with them—it is "distance … what distinguishes [a] loose and approximate grip from the complete grip that is proximity" (ibid., p. 261). By incorporating the car and the environment into their intentional arc, the agent can refine their grip on the world—they can distinguish which features of their world appear as *useful* and thus follow the course of action which is afforded by them, through utilising their perceptual and motor capacities alone.

This refined grip, along with the intuitive actions that come with it, are not representational entities. A tighter grip on an environment does not mean, as Metzinger may perhaps suggest, that one is refining one's internal representation of the world by adding more information into it. Rather it means that the agent learns how to *perceive* the environment in a way that is sympathetic to their expertise. No representations need to be invoked for the world is already *presented* to the agent, rather it is a matter of better perceiving what is already there. Similarly, the intuitive actions by which the expert responds to the world with do not have to rely on representation either. What to do and when is not stored representationally in the inner realm of an agent—they are instead captured in the disposition of an agent to act in a certain way. They belong not to representation, but rather form an element of the agent's intentional arc. They are a form of what Merleau-Ponty calls "knowledge in the hands" (ibid., p. 144): a dispositional habit "bred of familiarity" (ibid) by way of recognising—perceiving—one's past motor actions in the affordances offered to the agent by the world at present. Through practise, and the ever-refined perceptual link between ourselves and the world, we can act without ever needing to posit the existence of any mental representations. Even if one's faith in representationalism means that they did wish to posit their existence, it is not clear how they are to bear upon our actions. If our actions are rooted in our perceptions, and an always-tightening link

between ourselves and the world, it is hard to see how decoupled, offline representations could have any influence on this. The habitual knowledge in the hands on which our action rely "is forthcoming only when bodily effort is made, and cannot be formulated in detachment from that effort" (ibid), and so the mentally simulated *dry-run* actions that Metzinger proposes cannot tap into this habitual knowledge. Our actions therefore, must necessarily be *online,* called forth by the world presented, not the world *represented*.

4.8 WHERE ARE WE NOW?

We have seen that Metzinger offers no compelling explanation for his belief that our access to the world is one fundamentally mediated by internal representations; neither the speed at which the representation is updated, nor the principle of *metabolic price* offer good grounds to assent to his position. Moreover, Metzinger's assertion that representations are necessary for action has also been shown to be misguided—we can act successfully by relying on perception to inform our motor skills, rather than by positing a representational test-space where these actions are simulated before their execution. It seems therefore, that there is no motivation for accepting Metzinger's representationalism. As such, we can dismiss Metzinger's representationalism, and with this, Metzinger's assertion that the self has no ontological status beyond the illusory product of these representations.

4.9 AGAINST METZINGER III: METZINGER'S HIDDEN SELF

This said, showing Metzinger's entire representationalist project to be unmotivated is not the only way we can combat Metzinger's assertion that the self does not exist. It can also be done without calling the existence of representations into question. If we look at Metzinger's basic argument, we see that it is a variant of the *Virtual Self Theory* (VST

herein).[21] That is to say that Metzinger is committed to the position that "the brain represents the existence of a self with various properties [and] … the entity represented does not exist" (McClelland, 2017., p. 22). The self is merely intentional, with no real existence beyond our internal representations: "all that ever existed were conscious self-models that could not be recognised *as* models" (op cit., 2004., p. 1). Metzinger's view, as with all variants of VST, suffers from the fact that "denying the existence of the self is hard to reconcile with accepting the existence of self-representations, for [they] must surely have a bearer that can be appropriately described as the self" (op cit., 2017., p. 32). There are two ways in which one can leverage this reconciliatory difficulty against VST—either by arguing that VST does not rule out the possibility that "the self does exist but is systematically misrepresented by us" (ibid., p. 22) or by arguing that VST *implies* the existence of a self. We shall explore these two positions in turn, starting with the former.

McClelland argues that VST does not rule out the possibility that the bearer of the self-representations may be a self by appealing to the *content* of our self-representations. As we have seen, Metzinger suggests that the content of the PSM—our self-representation—is the ego. All of the features of our everyday understanding of selfhood exist as the *content* of our self-representation. This means that the PSM is a kind of *de se* representation; its content "refer[s] essentially to the representer of that representation" (ibid., p. 35). Therefore, it seems to be the default position that whatever bears a PSM, as it is a self-representation, must be a self: "according to our self-representations, whatever entity is the bearer of those very representations is the self" (ibid., p. 36). It is up to Metzinger (and other VST proponents) to show us why this cannot be the case—the burden of proof has shifted from realists about the self to demonstrate why there *is* a self, to the VST which

[21] Dennett's assertion that the self is a centre of narrative gravity is also a variant of this theory.

must “give us reason to believe that the referent of our self-representations is not [the bearer of those representations]” (ibid). Until this is done, the possibility that a self exists remains viable.

Though this is sufficient to undermine the attractiveness of the VST, we can go further and argue that VST is incoherent as it actually demonstrates the existence of a self. Let us revisit the earlier *first-pass* argument made and then dismissed by McClelland. The argument is as follows:

1. If VST is true, then we each have mental self-representations.
2. All mental representations have a bearer.
3. The bearer of one’s mental representations is the self.
4. Therefore if VST is true then there are selves.
5. Therefore if VST is true then it is false. (ibid., p. 32).

McClelland dismisses the argument by appealing to Dennett’s notion of Gilbert, the fictional self created by “a computer that has been designed or programmed to write novels” (Dennett., 1992., n.p.). McClelland agrees with Dennett’s assertion that, Gilbert, the fictional self, is the product of a “invention process in which there aren’t any selves at all” (ibid)—the machine, though supposedly creating a self, can hardly be thought of as a self in its own right—and draws the analogy between the self-creating novel-writing machine and the brain. Like the self-creating, novel-writing machine, the brain acts as “the originator … but not the *referent* of … self-representations” (op cit., 2017., p. 33). The referent of the self-representations is not the “real concrete entity that generates those representations” (ibid), but rather some other, “merely intentional entity” (ibid) and thus,

the proponent of VST is free to argue that premise three of the above argument is erroneous, and self-representations "do not require the existence of a self" (ibid) to bear them.

McClelland concedes the ground to Dennett and VST prematurely—in reality, Gilbert, the so-called self that is generated by the novel writing machine, is *not* the product of a self-less process and thus should not serve as evidence to dismiss the above argument. Let us agree with Dennett's suggestion that Gilbert exists as an intentional entity: Gilbert only exists insofar as he represented in the words and phrases in a machine-produced novel. But this is not the whole story. Gilbert cannot be reduced to the words of the novel themselves, for that would mean that in each closed book dwells a fully-fledged self. Rather, Gilbert's existence must rely upon the *meaning* of the words in the novel itself. As meaning is not a necessary condition of a word—there is nothing about the inscription 'chair', in and of itself, that suggests an object that is suitable for sitting—Gilbert's existence as an intentional entity *presupposes* the existence of an entity that can interpret the words that gives rise to Gilbert's (intentional) existence. Or, to put it in a different way, Gilbert's fictional life may be represented by the novel, but calling Gilbert a self (albeit a fictional one) is a *judgement* made by the person reading the novel: Gilbert's selfhood is not borne by the novel but is borne by the person *reading* the novel. In much the same way that there is no such thing as noon if there is nobody to be there to observe it, there is no Gilbert if there is no reader—the novel gives us Gilbert *qua* fictional entity, but it is the reader that makes that fictional entity into a self (in Dennett's sense). Gilbert's self-hood *relies* on the existence of a reader—a self—in order to recognise and confer the status of (fictional) self to Gilbert, for without the reader, there is just a pile of written-on paper: "something else [a self] has to gather up the successive states of the machine" (Tallis, 2020., p. 150) and

decide that there is a (fictional) self in there somewhere. As such, because "Gilbert's fictional self parasitizes the consciousness of readers who are real selves" (ibid), Dennett's suggestion that Gilbert is a product of a self-less process is wrong—without the existence of a self to confer selfhood to Gilbert, all the machine creates is a stack of paper: an object, not a self. If there is even a self there at all—which is itself unlikely if we consider Searle's assertion that "syntax is not intrinsic to physics [and] is always relative to an … observer who treats certain physical phenomena as syntactical" (Searle, 1992., p. 208)— then the inscriptions on the paper no more contain a self than the configuration of stones on a pebbly beach does. We can only say that they do because there was another self there to say so.

Because of this, McClelland's assumption that premise three of his argument fails is not justified by the evidence he presents—it seems that the existence of a self is a key part of *creating* a self (Tallis., 2020, p. 152). The novel alone is not enough to create a self; another self must be posited to give life, so to speak, to Gilbert. Dennett's argument therefore has no bearing on the argument presented by McClelland and in the absence of any other objections, there is no good reason not to accept McClelland's argument as sound. If the existence of a real self is key to the construction of a virtual self, then VST is self-undermining and so can be dismissed.

4.10 WHY MATERIALISM STRUGGLES WITH THE SELF

In this section, we have looked at three different ways that materialists might try to account for the existence of the self—first Van Inwagen, who suggested that the self was numerically identical with the brain, or one of its parts, then Dennett who argues that the self does not exist and is simply a narrative construction, and finally Metzinger, who argues that the self is again non-existent as it is the content of a representation with no

referent. All three of these strategies are bound together, not just by their materialism, but also by a faith in neuroscience to uncover the truth about selfhood. To explain things in a layman's fashion, Van Inwagen takes the implicit idea behind some understandings of neuroscience—that we are our brains—and makes the idea literal, arguing that so long as our brain survives, then so do we. Dennett and Metzinger also use neuroscience to motivate their claims. Moving away from Van Inwagen's idea on the grounds that it is a category error, they see that there is nothing that is uncovered by neuroscience that is a good candidate for selfhood so they switch to their illusionist positions. Neuroscience reveals the truth, so the illusion must be explained in terms that are inoffensive to neuroscience—language and neurobiological representations respectively. But what all three philosophers fail to note is that it is impossible to offer a materialistic explanation for the self because materialism is not the right tool for the job.

The reason for this should be familiar: it is the same reason why materialism struggles to explain consciousness and free will. Materialism operates from an Archimedean point—their view of the world is an objective view from nowhere, achieved via an explicit attempt to eliminate subjectivity from the situation. Materialism and its understanding of the world is necessarily self-less—the self is excised from enquiry to prevent any subjective aberration to the 'truths' that it endeavours to uncover. As such, it is hard to understand why Van Inwagen, Dennett, Metzinger, and those sympathetic to their positions believe that materialism is the appropriate tool to explain self-hood. By design, this is an impossible task—one cannot expect to find a self in a theory in which, for the sake of objective neutrality, it has been deliberately removed. Any attempt to do so must either identify the self with some material entity, or it must deny the existence of a self and try to explain the illusion of the self in materially inoffensive terms—but as we have seen above,

neither of these options are viable. If we are to have an explanation of the self, it cannot be one rooted in materialism—"personal identity cannot be translated into, or reduced to, 'it is'" (Tallis, 2020., p. 175). Selfhood, like consciousness and free will, belongs to that which materialism is unequipped to reveal.

5 THE SIGNIFICANCE OF THE IMPOVERISHED IMAGE AND THE DIFFICULTIES OF FORMULATING MATERIALISM

So far we have seen three ways in which the materialist image of humanity used to guide the development of maximally humanlike technologies is impoverished—the methodological and materialist commitments of the image cause it to neglect not only consciousness (chapter 2), but also human free will (chapter 3) and our sense of selfhood (chapter 4). By viewing humanity from an objective, third-person Archimedean point we become unable to properly grapple with the features of human subjectivity and so we are forced into denying the existence of these features and proclaiming their illusory nature. This limits the way in which our technologies can be like us. Features neglected in the guiding image cause these same issues to be neglected in the technological products informed by this image. Omissions in the blueprint suggest omissions in the product—thus giving us no reason to think that such technologies would be in possession of subjective faculties. With this in mind, should we recall the distinction between strong-AI and weak-AI from §1.2, it becomes clear that the impoverished image effectively renders strong-AI—technologies that possess a mind—impossible in the materialist paradigm. It is nonsensical to suggest that technologies, developed in line with an image of humanity that neglects subjectivity, will be in possession of subjectivity. As such, at this point, it seems that, when guided by the impoverished image, only weak-AI—technologies that are developed to *emulate* human faculties—is possible.

This said, there will be those, motivated by both faith in technological advancement and a belief that predicting the future is foolish, who are keen to overlook this issue as temporary and contingent: some future theory with a less impoverished image of humanity may one day arise to guide the development of humanlike technologies. This chapter seeks to shore

up our conclusion that strong-AI is impossible in the materialist paradigm by demonstrating that the materialism we have inherited from the prevailing doxa is unfit as a justification for the belief that a technology which *appears* to be in possession of subjective faculties is *actually* in possession of such faculties. To do this, we first introduce a thought experiment whereby a person, unbeknownst to them, interacts with such a technology: the synthetic stranger. We then seek to understand *why* they would believe that this stranger is in possession of subjective faculties— we first explore a response motivated by a naïve appeal to the stranger's behaviour, before then exploring a response motivated by materialism. Following this, we then examine the difficulties in formulating materialism, in order to demonstrate further that relying on materialism is not motivated by rationality: we only do so because we assume materialism to be more robust than it is, because we have inherited it from the prevailing doxa.

5.1 MEETING THE SYNTHETIC STRANGER

Imagine you are one of your future descendants; you read the newspaper. It says that scientists have created a strong-AI and will be exhibiting it in a nearby city. You travel to the exhibition and wait in line to meet the strong-AI—a steward ushers you into a waiting room and in this room, there is another person. You sit down and exchange pleasantries with the stranger—they make small talk before getting a cup of coffee from a nearby vending machine. They return to their seat and take a sip of their coffee and swear under their breath as the hot coffee burns their tongue. Your name is called and you bid farewell to the stranger in the room with you. The next room is empty apart from a woman in a white lab coat, sitting at a desk—she asks you to sit and reveals that the stranger in the waiting room was actually the synthetic agent you came to see. You express your surprise and she asks you a series of questions about your experience of the synthetic agent. One question is: *in your opinion, is the stranger in the previous room conscious?*

There are a number of different answers that one can give to this question but let us assume that the scientist's questionnaire only affords a binary yes-or-no answer. We shall focus our attention to those that answer affirmatively—those who think the stranger *is* conscious—for it is this group that we seek to dissuade. Now, let us assume that the scientist is not just collecting quantitative data, but also seeks to understand the rationale of the people that answer the questionnaire: under the section where one gives their response, there is room to justify their answer. What could we expect to see in this section?

5.2 YES, THE STRANGER IS CONSCIOUS: BEHAVIOURISM

The most obvious justification appeals to a kind of naïve behaviourism, after all, we have little other information beyond the way they act to justify our opinion. We cannot pass them through any kind of imaging equipment, nor could we dismantle them to see what it is they are made of, and how they are put together. The only way that we can investigate the stranger is by way of observing their behaviour: it is natural that this should then form the basis of our justification. Of course, there was nothing suspicious about the stranger's behaviour to cause us to doubt their consciousness. They did not delay their responses in order to give them time to process what we had said, they did not offer nonsensical replies or do anything else to rouse our suspicions. So we arrive at the position that the stranger is conscious—their actions were consistent with what we expect from conscious agents, so we think that they are a conscious agent too.

Such reasoning enjoys a significant amount of layman plausibility—if nothing presents itself to distinguish the stranger from a conscious agent, then the agent must be conscious; for instance, when the stranger went to get their coffee, this seems to betray that the stranger had a conscious desire for the coffee and the free will to get it, and similarly, when they burnt their tongue on the hot liquid, this seems to betray that the stranger was

phenomenally conscious of the pain caused by burning their tongue. Humanlike behaviours thus suggest that whatever acts in a human way is in possession of a humanlike mental life. Such rationale seems to haunt the common-sense criteria by which we come to ascertain the consciousness (or lack thereof) of a humanlike technology. For example, the Turing test seems to share this naïve behaviourism—that if something *acts* as though it is human, it *is* in some way human—as its impetus. The test, at its heart, is one of mimicry: when distilled into its most salient parts, the Turing test suggests that if a machine can act in such a way that we cannot reliably distinguish it from a human, then we can say that the machine is exhibiting intelligent behaviour (Turing, 1950., pp. 433 – 435). But this, as all such appeals to behaviour, is problematic.

Let us consider an argument against the Turing test that elucidates its problematic nature: Searle's Chinese room argument. Imagine someone ignorant of the Chinese language is locked in a room with a large stack of papers written entirely in Chinese. The characters on the papers present themselves to the person as nothing more than "meaningless squiggles" (Searle, 1980., p. 418). The person is then given a second stack of papers with a set of rules, written in English (a language in which the person is completely proficient) on how to link the first stack of papers with the second stack. The person, equipped with these rules can match one set of Chinese characters with another, simply by comparing their shapes. Now imagine the person receives a third set of Chinese papers, and a second set of English instructions—these instructions allow the person to sort the Chinese papers in such a way that they can duplicate some of the Chinese characters and give them back to the person supplying the instructions. With practise, the person in the room gets particularly proficient at this exercise and returns reams of Chinese characters at a great pace but still with absolutely no understanding of what the Chinese characters signify. The characters

are only significant in the context of their shapes and the English instructions. From the perspective of someone fluent in Chinese, the situation appears to be thus: a text is given to the person in the room (the first stack of papers), and a series of questions are then asked about the text (the second stack of papers), to which a set of sensible and relevant answers are returned: "nobody looking at [the] answers can tell [the person in the room doesn't] speak a word of Chinese" (ibid). If one were to ask a person fluent in Chinese if the person in the room was also fluent in Chinese, they would undoubtedly answer that they were—but this is not the case. From the perspective of the person fluent in Chinese, it is a sensible assumption to make, but from the perspective of the person in the room, such an assumption is erroneous. The person in the room is doing nothing other than using the rules that are written in a language that they can understand to manipulate the texts that are written in a language which they cannot—Chinese is just as opaque to the person in the room as it was when they were first locked in there. In short, the action of supplying sensible answers in Chinese to questions written in Chinese is not a reliable indicator that the person supplying the answers can understand Chinese.

This also holds true for the Turing test: a machine able to act in a way that suggests that the machine is intelligent presents no reason to believe that the machine can actually grapple with the concepts, form its own understandings of phenomena presented to it, or engage with anything else traditionally associated with intelligence. Similarly, this rationale can also be extended to the synthetic stranger in the waiting room—just because they *seem* to be exhibiting conscious behaviours does not mean that they are properly in possession of consciousness. Just as the question-answering efficacies of the person in the room are not a reliable indicator as to whether they understand Chinese, the humanlike behaviours of a synthetic agent are not an infallible indicator of a humanlike mental life. Concepts

associated with mentality (understanding, intelligence, consciousness, etc.) stand distinct from behaviour. It is possible for an agent to *act* as if they understand, as if they are conscious, or as though they possess intelligence, and so on, without this being the case, and so behaviour is no infallible indicator of the possession of these faculties. Those that justify their claim that the synthetic stranger is conscious by appealing to their behaviour are guilty of not recognising the fallibility of such appeals. This is perhaps forgivable, if the respondent were to meet the stranger in less unusual circumstances, but the questions of the researcher should be sufficient for the respondent to re-evaluate the value of using behaviour alone to motivate their assumptions.

5.3 YES, THE STRANGER IS CONSCIOUS: MATERIALISM

Let us assume that some of those that went to see the stranger did so because they have been following the development of the stranger—they are somewhat *au fait* with the project and the ideas that motivate it. Perhaps they even are aware that behaviour is no infallible indicator of consciousness. In this case, what would motivate their claims that the synthetic stranger is conscious? In lieu of behaviour, we can suggest that their claims are motivated by the background belief in materialism that is given to them by the prevailing doxa. As we saw in chapter 1, the idea that all phenomena—including human subjective faculties—are either material in their own right, or somehow supervene on the material is widespread by virtue of its hooking into the Promethean attitude. If materialism is true, then all there is—in theory—to making the stranger replete with human capacities, is the arrangement of matter to create certain interactions to give rise to the capacities: holding materialism to be true makes the idea of creating a technology in our image logically possible.

We shall now take something of a detour—we will explore the difficulties in giving a reasonable definition of what it means to be a materialist, before then exploring Hempel's dilemma. Once this is done, we shall return to our respondents to demonstrate that it is not rational to let the prevailing materialist doxa inform their belief that the stranger is in possession of subjective faculties.

5.4 MATERIALISM: DIFFICULTY IN DEFINITION AND VAGUENESS IN CONTENT

Using materialism to justify a claim that the stranger is conscious is problematic, not at least because it is unclear that a reasonable definition of materialism can be offered. The most basic understanding of materialism "is the thesis that everything is physical" (Stoljar, 2010., p. 28). Indeed, this is perhaps the most likely formulation to be endorsed in a layman context, but such a formulation is problematic as it cannot account for *abstract* particulars. For example, it does not seem correct to say that the UK Government is a *material* entity or that there is a particular collection of matter that is self-identical to the UK Government, if for little else for the fact that the Members of Parliament that form the UK Government are shifting—the Chancellor of the Exchequer may be one collection of matter one day, and another the next. Instead, it makes more sense to assert that the UK Government is a legal, political or legislative body, rather than a purely material one (ibid., pp. 29 – 30). Such an argument prompts the abandonment of this naïve version of materialism; two possible amendments to the above formulation can be posited.

Firstly, perhaps it seems that we can escape the above objection by suggesting that materialism is true if "every concrete particular is physical" (ibid., p. 31), but such a formulation opens the door to property dualism—yes, it may be so that all concrete particulars are material, but this does not preclude them from having immaterial properties (ibid., p. 32). This issue can be avoided by amending the original formulation of

materialism in such a way that materialism is true when every instantiated *property* is material (ibid., p. 36), i.e., if properties such as roundness, greenness, etc., can be owed to matter then materialism is true. This again runs into issues. Not every property is material (ibid., p. 33)—for example, the number twelve has the property of being even, being divisible by three and being the square root of 144, yet these are not material properties, rather they are *mathematical* properties. As such, we must amend the statement further—we must claim that all instantiated properties are either material properties, in and of themselves, or properties that are *necessitated* by an instantiated material property, i.e., if material property X is instantiated then legal/social/mathematical/etc., property Y is also necessarily instantiated alongside it (ibid., pp. 37 – 38). This formulation of materialism is liberal enough to not fall foul of abstract entities, but it not so liberal as to allow the conceptual space needed for property dualism to cause any issues—as such, it seems a reasonable candidate for a formulation of materialism, although, this formulation does not yet seem to be complete. If it is so that all properties are either material in their own right or are *necessitated* by material properties then we must make clear what just is meant when we refer to material properties.

One simple understanding of the notion of a *property* is: "a property is the way an object is" (Daly, 1998., p. 196). Should a book be safely stored away on our bookshelf, the book has the property of being on the bookshelf—as such it can be assumed that material properties constitute a subsection of the totality of ways an object can possibly be. If we were to provide an exhaustive list of ways an object can be we would be able to sort the entries in the list into categories—one may be chemical, another legal, a third material and so on.

This suggests the existence of some criteria we can use to distinguish one category from another (ibid., p. 197)—but what are the conditions a property must meet for it to qualify as material? Stoljar posits that a property is a material property if it is one of the distinctive properties of material objects, that is also: predicated in a physical theory, objective, knowable through scientific methodology and is *not* one of the distinctive properties of some immaterial entity, e.g., the soul (op cit., 2010., p. 57). Such criteria are reasonable and speak to the common-sense understanding of what it means for a property to be material. Should we then combine this understanding of material properties with the above formulation of materialism we are led to a new formulation of materialism: materialism is true if all instantiated properties either meet the criteria for a material property themselves, or are necessitated by an instantiated property that meets the criteria for a material property. Should this formulation be adequate, then our respondent would—aside from the issues explored in the previous chapters—be justified in using materialism to justify their claim that the synthetic stranger is conscious: the respondent could argue that the stranger possesses the property of *being-conscious*, for it is necessitated by the material properties also possessed by the stranger. Following Stoljar's terminology, we shall call this formulation of materialism the starting point view (SPV herein).

In order to evaluate the adequacy of the SPV, let us assume that we have three possible worlds. In each a different metaphysical theory is true—in one world, the classical atomist theory as articulated by Democritus is true, in another a Newtonian, *atomism-plus-gravity* theory is true and in the third, some consistent subset of our modern physical theories are true. Two questions should be asked: "is materialism true in this possible world?" and "does the SPV hold in this possible world?" Should the answers to these questions agree in the context of a particular possible world—i.e., both answers turn up false, or true for the

given possible world—then it seems logical to suggest that the SPV is a reasonable formulation of materialism (ibid., p. 58).

Classical atomism holds that the world is constituted of supremely minute particles: if one takes an object and splits it in half, and then take one of those halves and split it in half again, and so on, eventually one will reach a stage where the half cannot be split further—this minute entity is the atomic unit that makes up the original object (Kenny, 2007., p. 27). The atoms of classical atomism can, therefore, be thought of as being "rather like rocks only much, much smaller [as] like rocks, atoms have the properties that intuitively physical objects have" (op cit., 2010., p. 59). These properties are the standard properties associated with material objects: location, size, shape, solidity, etc. Classical atomism also holds that worldly phenomena are explainable by way of reference to these atomistic units. Now should the questions above be asked in the context of this world, it is clear that both answers would be yes—not only is materialism intuitively true but also, the SPV holds (ibid., p. 58). SPV thus passes this test.

The Newtonian world bears a great degree of similarity to the world of classical atomism—the only distinction is the addition of gravity. Whereas in the classical atomist world, atoms only interacted with each other directly, i.e., by literally colliding with each other, in the Newtonian world, atoms can also interact with each other indirectly, via gravity: one atom may attract another without ever actually coming into contact with it (ibid., p. 60). As with the classical atomist world, it seems obvious that materialism is true in this world, but whether one thinks the SPV holds depends on how one thinks of gravity—should one think gravity to be a property of material objects, then the SPV holds,

should one not think that gravity is a property of material objects, then the SPV does not (ibid., p. 62).

It is difficult to see how gravity could be the property of a material object—it seems more intuitive to argue that gravity is not a property of an object, but rather is a phenomenon that exists purely in a relational form between two (or more) objects. Should one consider a profoundly empty world consisting of one single atomistic unit suspended in the void, then the claim that gravity exists in this world is intuitively implausible. The atom would still have a location, a size, a shape and so on, but it does not seem clear how it could have any *actually instantiated* gravitational properties—the atom is neither attracting anything, nor is it being attracted *by* anything. Yet, if gravity *were* the property of a material object, then we would be forced into concluding that gravity exists in this world. As such, this undermines the attractiveness of the SPV considerably as it forces us into picking a side on this debate. Two people could agree that the world is made up of atoms and gravity, but if one believes gravity to be a *relational* property then, under the SPV, they are forced into denying that their world is materialist: such contingency and looseness is unwelcome in a formulation of materialism.

Our evaluation of the SPV is not yet complete—before we can rule it out as a formulation of materialism, we must first see how it fares in the modern physics world: if it is as serviceable here as it was in the classical atomist's world, then such a formulation may still be robust enough to be useful. Let us look at two versions of the modern physics world—one where the theory of relativity is true and another where quantum theory is true.[22] In

[22] Of course, it is possible for *both* theories to be true in the same world, but here I separate the two theories out in order to better demonstrate my point.

both versions of the modern physics world, we are told something that undermines the SPV by way of calling into question the intuitive properties of material objects. In our intuitive understanding of the world, we commonly assume that the size and shape of a material object are constant, *ceteris paribus*. So long as the particular object remains as it was—i.e., nothing is removed or added to it—it has a particular size and a particular shape. We also intuitively think that the size and shape of the object would remain constant in all situations; a kitchen chair will be the same size should we take it upstairs, or should it be in the back of a removal van travelling along the motorway. Common-sense understanding tells us an object has a constant size and a constant shape; relativity claims this understanding is incorrect. In the world where relativity is true, the objects that we believe to be in possession of a constant size and shape instead have a size and shape that can fluctuate in relation to the speed at which an object moves. In this world, the chair in the back of the removal van will be squashed and condensed in comparison to the same chair sitting still in the kitchen (Stannard, 2008., pp. 13 – 16), At such low speeds, the compression of the chair will be minimal, but the theory of relativity argues that this contraction "applies at whatever speed [the object] travels" (ibid., p. 15) with the contraction becoming more pronounced as one approaches the speed of light—"right up close to the speed of light, [the object] could be flattened thinner than a CD" (ibid).

Now, in the context of this version of the modern physics world, what are the answers to our questions? Is this a world where the theory of relativity can hold true purely utilising material facts or properties? It seems plausible—no reference to anything immaterial is necessary for relativity to be true. But is the world materialist? It does not *necessarily* seem as though it is: nothing exists in the theory of relativity that would conflict with a dualist's view of things. The dualist can agree with all of the unintuitive implications of the

relativity theory—time dilation, distance contraction and so on—and still maintain that humans possess some manner of immaterial mind or soul. Endorsing one does not mean we have to abandon the other. But does SPV hold in this world? It does not seem so. The SPV asserts that in order for a property to be a *material* property, it must, in part "be one of the distinctive properties of physical objects" (op cit., 2010., p. 57)—it is difficult to see how a size and shape that alters when the object is in motion is in line with this assertion. Having the property of *gets compressed when in motion* is *not* something that we would intuitively apply to objects in the world, as it runs counter to the relative permanence that we would usually attribute to objects and so whilst materialism is true here, the SPV does not hold.

Let us change our focus to the version of the modern physics world where quantum theory is true to see how SPV fares there. Again, in such a world, we are told something unintuitive that runs counter to the SPV. In the quantum theory world, it is true that there are particles. It is also true, that these particles exhibit wavelike properties (Polkinghorne, 2002., pp. 18 – 20). Consider the famous double slit experiment: should one pass monochromatic light—light of one particular wavelength—through a shield with two small slits in it, then the light that passes through the slits will, when hitting a screen beyond the slits, form an interference pattern. When two peaks in the waves meet, the light is amplified and a bright section is produced. When a peak in one wave meets a trough in another, the waves will cancel each other out, producing a dark section. These dark and bright sections alternate, forming the interference pattern. Now, should one repeat this experiment with particulate matter, one might expect that two collections of matter would gather in relation to the slits in the shield—for example, if we were to pass sand through the slits, we may expect two piles to form beyond the slits—but this is not so. In the

possible world where quantum theory is true, should one fire an electron—an example of particulate matter—through the two slits and onto a screen that records where the electron lands, an interference pattern emerges thus demonstrating the wavelike behaviour of the electron (ibid., pp. 22 – 25).[23] Should we subject this quantum theory version of the modern physics world to our questions, it is again obvious that the quantum theory world relies on no immaterial properties or facts in order to be true, but does the SPV hold? It does not seem likely. The things that quantum theory tells us about "are not intuitively physical objects, and do not have the properties of intuitively physical objects" (op cit., 2010., p. 65). It is not intuitive that the material properties of objects in the world are necessitated by entities that are not intuitively material—i.e., entities that exhibit both wavelike and particle-like behaviours. So in both of our versions of the modern physics world (in addition to the Newtonian *atomism-plus-gravity* world), it is so that materialism may be true (insofar as both *can* hold true without the need of immaterial entities), but SPV does not hold, thus the SPV—the formulation of materialism that suggests materialism is true if all properties are either properties that meet the criteria for material properties explained above, or necessitated by properties that meet the criteria for material properties—has been defeated on the grounds that there are possible worlds where materialism is true, but the SPV does not hold.

So far we have seen that the SPV only seems to be a sufficient formulation of materialism in the world of classical atomism: in the classical atomism plus gravity world, and the two versions of the modern physics world we have seen, the SPV does not hold. The weakness of the SPV is self-evident. It is too strong a formulation, and thus forces us to deny

[23] Should the electron be *observed* passing through the slits, unusually, the interference pattern does not form as the electron acts how we would intuitively assume it would act—it only appears to act as a wave in the absence of observers. This detail is left out in the discussion above, for the sake of simplicity.

materialism in worlds we would intuitively believe to be materialist, and so we can dismiss it. In trying to account for a reasonable formulation of materialism, it may be fruitful to return to our understanding of what a material property is. If we can supply an understanding of material properties that is more liberal than that SPV employs, we may be able to offer a less problematic formulation of materialism. There are a number of ways that we can liberalise our understanding of material properties: we shall look at three such ways in the following section.

5.5 THREE OPTIONS FOR LIBERALISING MATERIAL PROPERTIES

This section explores three options for liberalising our understanding of material properties. Firstly, I look at Christopher Daly's family resemblance account of material properties, before dismissing it as flawed. Secondly, I look at Paul F. Snowdon's natural kind account of material properties before again dismissing it as flawed. Finally, I argue that David Papineau's suggestion that material properties are those predicated in a physical theory presents the most attractive formulation of material properties explored here.

The family resemblance account of material properties holds that a property is a material property insofar as it bears some resemblance to one (or more) of the paradigmatic material properties in the world (Daly, 1998., pp. 200 – 203). Such an account assumes that there are a set of paradigm defining properties that are thought to be material—e.g., perhaps the traditional Lockean primary qualities of size, shape, location, etc.—and we can discern if another property is material or not, on the basis of whether it resembles one of these paradigm properties. Consider a pen—although there is maybe a sense in which the pen is not material, *per se* (at least in the sense that 'pen-hood' is a particular interpretation of a particular collection of matter) there is also a sense that the pen is a purely material object. This is due to the fact that the pen, by virtue of its constituents, *resembles* the paradigmatic

properties of material objects. The pen has a size, a shape, and so on. Now consider the property of 'being Prime Minister of the United Kingdom'; this is an immaterial property for it does not resemble any paradigmatic material properties. Though the person instantiating the property does indeed have a size, a shape and so on, 'being the Prime Minister' itself does not, and therefore cannot be through of as a material property. Though plausible, this account is also flawed.

Assume there is a world where a whole set of properties resemble each other, in an incremental fashion. Property A resembles property B, which resembles property C, and so on. Eventually, at some point along this line of incremental changes, a property will exist that bears no resemblance to the earlier properties. Should we compare property A and B, they will be similar but should we compare property A with property Z, then there would appear to be no resemblance between these. In a world where the entire sequence of properties, from A to Z, are existent, it may be argued that this poses no issue: Z counts as a material property because it resembles Y, not because it resembles A, but this is not so. Let us say that property A, at the beginning of the chain, is by definition, material and let us say that property Z, at the end of the chain, is by definition, *not* material (perhaps it is a legal property, or something similar). Those committed to the family resemblance account would argue that this is mistaken: there cannot be a resemblance between properties A and Z, for one is material, and the other is not. But this is not so—both material property A and immaterial property Z can resemble each other on the grounds that they are both currently instantiated or that both are existent in time. Insofar as this is possible, the family resemblance account of material properties cannot hold, because there is the potential for material properties to resemble immaterial properties—as such we must abandon this family resemblance account of material properties (ibid).

Another option in liberalising our conception of physical properties is to posit that "a state of affairs is a physical one if it consists only of physical objects instantiating physical properties" (Snowdon, 1989., p. 154). As Snowdon claims, this notion suggests that the term 'physical' is "an extremely basic natural kind term, a term for the most all-embracing (natural) kind of which we are acquainted" (ibid., p. 153). This thesis suggests that material objects form a natural kind, a totality of objects that are replete with the same essential features, features that we can discover in an *a posteriori* fashion. To unpack this, a material property is one that pertains only to the common essence of the objects that form the natural kind of material objects (op cit., 1998., p. 203), as such, we can call this understanding of material properties the natural kind account. This account runs into three issues.

Firstly, an element of circularity exists in this formulation of material properties—if material properties are those that are essential to material objects, as Snowdon suggests, then we are still left with the question of what 'material' actually means. To respond to the question '*what are material properties?'* with '*material properties are the properties of material things'*—as the natural kind account does—is obviously unhelpful.

Secondly, the natural kind account does not rule out property dualism (ibid., pp. 205 – 206). The natural kind account holds that material properties are those that would be present in a world solely constituted of material objects—property dualism holds that it is possible that material objects can possess non-material properties (for instance, a material person may have non-material mental properties, e.g., the property of 'liking-tea'). As property dualism is possible in a world totally constituted of material objects, the natural kind account is flawed.

Thirdly, even if we are to dismiss humans from the possible all-material world, there are number of properties held by classically material objects that we would not necessarily wish to describe as material. Chairs would be present in an all-material world, but if the chair was wooden, then it also possesses *biological* properties by virtue of the fact that its wood is made up of cells (ibid, p. 205). With the natural kind account, it follows that there are no distinctions between kinds of properties: all properties instantiated in an all-material world would necessarily be material under such an account, but the existence of *biological* properties (amongst others) shows this to be misguided. The circularity that seems to haunt the natural kind account, in addition to the fact that it cannot guard against property dualism, and its overly liberal labelling of properties as material are all good reasons to dismiss this attempt at liberalising our understanding of materialism.

To understand the third option we have available to us in liberalising our understanding of material properties, we should note that the family resemblance account and the natural kind account though flawed, do both contain within them a common and valuable element for another account of material properties—they both seem to speak to the natural sciences in some way. Whilst the family resemblance account is not of much use when formally defining material properties, T. S. Kuhn (1977, pp. 309 – 313) suggests the notion of family resemblances is one that is fundamental to the investigative practices typical of science (at least in a naïve, proto-scientific sense). In such practises, one does not often operate with formal and rigid definitions or concepts, but rather one observes the phenomena and pares specific elements of our experience off into groups by virtue of the similarities (or lack thereof) in their qualities: we learn "to apply symbolic labels to nature without anything like definitions or correspondence rules … [but rather using] primitive perception[s] of similarity and difference" (Kuhn, 1977., p. 312). Similarly, the natural

kind account pays homage to the natural sciences by asserting that the properties of natural kinds are discoverable via empirical experiential means. This gestures towards an alternative account of material properties—an account that claims that material properties are "whatever categories [that] are needed to give full explanations for all physical effects" (Papineau, 1993., p. 30), i.e., that they belong to a physical theory that can explain the goings on in the world, without invoking the immaterial, or without "mak[ing] use of any psychological categories" (ibid). What this means for our respondent is that their claim that the synthetic stranger is conscious is justified, so long as one can, in principle, explain the stranger's purported consciousness by relying entirely upon facts and concepts that feature in a physical theory.

Following the terminology of Stoljar (2010., p. 69ff) we shall call this account of material properties the theory view. As suggested by Papineau, such a view holds that a property is a material property on the basis that it finds itself predicated in a physical theory, therefore, materialism proper is true if every instantiated property is either predicated in a physical theory itself, or is necessitated by those properties that are predicated in our physical theories. The appeal to theory found within this formulation prompts the question of what constitutes a theory. The best candidate for this—one that respects the general and abstract nature of materialism *qua* theory—is to understand a theory as the collection of ideas that physicists employ in their investigations of the objects in the world that we would intuitively hold to be material. Keeping the definition relatively open in this way means materialism can be true in multiple worlds—it is not, to its detriment, wedded to the peculiarities of the physical theories in one world or another. The theory view can be manifest in four forms.

5.6 THE THEORY VIEW I: THE ACTUALIST THEORY VIEW

To begin, we shall look at the two forms of the theory view that depend on whether the theory in question is true in the actual world, or only simply true in a possible world. The first of these—the *actualist* version of the theory view (ATV herein)—holds that a property counts as material "if and only if [it] is expressed by a physical theory that is true at the actual world" (op cit., 2010., p. 75). For example, let us assume that in the actual world, quantum theory is true under a particular interpretation. According to the ATV, the properties that are significant to quantum theory—e.g., wave-particle duality—would count as material properties and materialism would be true because all properties that were instantiated in the actual world, were either material properties themselves, or necessitated by the existent material properties. This understanding seems to present no issues—that is until one introduces the notion of *twin-physics* (ibid., pp. 77 – 78). Imagine another world identical to the actual world in all ways, except that another physical theory is true. All properties of the actual world are replicated, but it is just so that they are necessitated by the premises articulated by another physical theory. This *twin-physics* world is not one that requires the addition of immaterial entities into our ontology—it is as material as its kin—but according to ATV we would be forced into denying that this world is materialist. To ATV only the holding true of *one* physical theory counts as materialism—that which is true in the actual world. So, in line with our example, all worlds where quantum theory is true are materialist, and all worlds where *another* theory is true are not materialist, regardless of whether they can operate without invoking the immaterial or not. As such, the *twin-physics* thought experiment shows ATV to be untenable as a formulation of, materialism. Another formulation is therefore needed: one that does not force us into the conclusion that a materialist world cannot only be the product of *one* physical theory. We explore this below.

5.7 THE THEORY VIEW II: THE POSSIBILIST THEORY VIEW

There is a more liberal alternative to the ATV: the *possibilist* theory view (PTV herein). PTV holds that a property counts as material insofar as it "is expressed by a physical theory that is true at some possible world or another" (ibid., p. 75). This version of the theory view runs into its own issues. Consider the classical dualist world. Here there is particulate matter—much like the matter in the classical atomist world—and when it is combined in the proper arrangement it can possess properties traditionally thought of as immaterial, e.g., the capacity to think, the possession of a soul, etc., that are not epiphenomenal. It is clear that in this world, materialism cannot be true—the world is paradigmatically dualist—but similarly, it is also true that the PTV holds in this world as it operates with a fundamentally physical theory: it is just so that the consequences of the form and interactions of matter can cause the immaterial to arise. Insofar as this is true, PTV is obviously unattractive—a formulation of materialism that allows paradigmatically dualist worlds to be classed as materialist falls foul of being too liberal a formulation of to be useful.

5.8 THE THEORY VIEW III: THE CURRENT THEORY VIEW

Two further versions of the theory view can be articulated on the basis of whether the theory central to its specific version of materialism is a *current* theory, i.e., one that holds that "current scientific findings provide support for physicalism" (Melnyk, 2003., p.14) or an *ideal* theory, i.e., "one that identifies the physical with the posits of [a] complete, ideal physical theory" (Dowell, 2006., p. 26). We shall refer to these versions of the theory view as the current theory view (CTV herein) and the ideal theory view (ITV herein) respectively. The CTV relies on a certain level of epistemic optimism that "those theories that are the object of *consensus* among current physicists" (op cit., 2003., p. 15) are *true*. If our current physical theories provide sufficient support to materialism, then this is based

upon the claims that: (1) our current physical theories are "typically approximately true and [that] more recent theories are closer to the truth than older theories in the same domain" (Laudan, 1981., p. 19) and (2) that these theories pick out actually existing entities in our worlds, i.e., that "there are substances in the world that correspond to the ontologies presumed by our best theories" (ibid). Synthesising these claims suggests that our current physical theories pick out genuine features of our world and are the closest one can hope to get to the truth. This though, is a claim that cannot stand as a mere assertion—newness does not necessarily entail closeness to truth, nor does it necessarily entail success of reference—some justification must be given here. The above assertion is motivated by the idea that if a scientific claim is broadly true and also picks out genuine worldly features, then it will it also "typically be empirically successful" (ibid., p. 21). This is to say that if a scientific theory is "explanatorily useful" (ibid., p. 31) in the sense that it can explain or justify the goings-on in our experiences then it must both possess some manner of truth, (even if that truth is not necessarily exact) and refer to genuinely existing features of our world.

This becomes problematic when one explores some scientific history—it is possible to find multiple examples of scientific theories replete with explanatory usefulness that utilise concepts or terms that fail to refer to anything that actually exists, and thus cannot be true: "if there were nothing like genes, then a genetic theory, no matter how well confirmed it was, would not be approximately true" (ibid., p. 33). Examples of such non-referring theories include those that were, at one point in time, commonly held to be true: e.g., "the humoral theory of medicine; … the vibratory theory of heat; … the vital force theories of physiology; [and so on]" (ibid). This presents issue to CTV because it challenges the idea that an empirically successful theory is both genuinely referring and approximately true;

the explanatory usefulness of a theory does nothing to demonstrate that it is true, or that its concepts genuinely pick out features of our world. There is nothing standing in the way of the possibility that our current theories are like those now rejected past theories that possessed explanatory usefulness, but ultimately failed to refer to any actually existing entities—"[if] physicalist principles are based on current physics … there is every reason to think that they are false" (Hellman, 1985., p. 609). In light of the fact that we do not know that our current theories genuinely refer, we must abandon CTV and shift our attention to ITV.

5.9 THE THEORY VIEW IV: THE IDEAL THEORY VIEW

ITV operates on the basis of an *ideal* physical theory (op cit., 2010., p. 94) . If we imagine scientific endeavours as an ongoing process, and grant scientists their epistemic optimism, it is sensible to suggest that these endeavours will eventually reach a point when they are finished, i.e., when science has uncovered all there is to be uncovered and formulated a true theory that suitably and successfully explains everything that falls under its remit. This *finished* theory, is an *ideal* theory. An ideal theory of genetics will be a complete and comprehensive explanation of genes, their roles, their constitution, and so on. An ideal physical theory will, likewise, be a complete and comprehensive explanations of physics. An ideal theory, therefore, is characterised by its completeness and its truth. ITV argues that a property counts as a material property if it is one that plays a part in an ideal physical theory, i.e., a true, complete theory of physics. This makes the ITV formulation of materialism one where all instantiated properties are either expressed by an ideal physical theory or are necessitated by properties expressed by an ideal physical theory.

Prima facie, this is compelling, though it encounters a fatal objection. If our current theory is not an ideal theory, which is certainly so, then we do not have any idea of what this ideal

theory entails, and thus a commitment to the ITV formulation of materialism is a leap into the dark. The unattractiveness of this is obvious—it seems ill-advised for the materialist to offer assent to a theory that has, in essence, *carte blanche* concerning its content by virtue of its obscurity: such logic is not dissimilar to buying a house before one has seen it. For instance, it may be so that a modified version of a classicalist dualist position—a physical theory that broadly claims that the immaterial can arise from the arrangement and mechanics of matter—is representative of the ideal physical theory which we may one day arrive at. If so, it is clear why the materialist would not wish to offer their assent to this theory; it may lead them into endorsing a physical theory incompatible with their materialist convictions, which consequently invalidates their motivations for assenting to the ITV formulation of materialism in the first place. As such, the ITV formulation of materialism goes the same way as its predecessors and must be abandoned.

5.10 HEMPEL'S DILEMMA

Combining the criticisms levelled at CTV and ITV leads us to an argument that undermines the motivation for assenting to a materialist metaphysic—*Hempel's Dilemma* (Hempel, 1980., pp. 194 – 195). Hempel's dilemma is clear: it is either so that materialism operates using a physical theory that is currently held to be true, and thus is false "since it will no doubt undergo further changes" (ibid., p. 194), or it operates using an ideal physical theory, in which case, "it is quite unclear what is to be understood … by a physical phenomenon" (ibid). Therefore, according to the dilemma, materialism is either false, or we have no idea what it means to be a materialist. This leaves the materialist in a difficult position. Due to the lack of a solid understanding of what materialism actually means, calling oneself a materialist is not an indication of their assent to a particular formulation of materialism—no satisfactory formulation currently exists—but is rather indicative of a conclusion prematurely drawn. Perhaps once an ideal physical theory has been articulated

(if such a thing is possible) then calling oneself a materialist, of justifying one's beliefs through reference to the material may cease to be premature, but until then, to call oneself a materialist is to endorse a false, or unknowable metaphysic.

5.11 BACK TO THE SYNTHETIC STRANGER

With the difficulties of formulating materialism made clear we can now return from our detour to our thought experiment. It is obvious that the respondent who uses the materialism inherited from the prevailing doxa to ground their claims that the synthetic stranger *is* in possession of subjective faculties is *not* justified in doing so. By letting the doxa inform their metaphysical position, they are ignorant of the difficulties in attributing meaning to the term 'materialism' and they are unaware that their common-sense adherence to materialism entraps them into endorsing a metaphysic that is ultimately—as Hempel's dilemma tells us—either false or unknowable. Such difficulties are effectively obscured by the Promethean attitude: because materialism does not gain assent through its robustness or argumentative panache, but rather gains assent by way of parasitising the human wish for a greater quality of life, the issues associated with the position are masked. In those laymen with a rudimentary grasp on materialism, we see that their claims regarding the stranger's subjective faculties are motivated by a tacit faith that the metaphysic lent to them by the doxa is true. In those who have grappled with the intricacies of materialism and yet still regard the metaphysic to be serviceable, this is motivated not strictly by a belief in materialism *qua* materialism, but rather because endorsing materialism hooks into some other, greater purpose that purports to eclipse materialism's flaws: the drive towards a better quality of life.

With this in mind, our assertion that strong-AI—i.e., technology in possession of a mind—is impossible under a materialist paradigm is shored up. Not only is the image of humanity

that guides the development of such technologies impoverished, insofar as it neglects our consciousness, our free will and our selfhood, but even if such a technology was supposedly developed, we could not justify the claim that it is in possession of subjective faculties. Attempts to do so on the basis of observed behaviour are undermined by the fact that behaviour is a fallible indicator of the possession of subjective capacities. Attempts to do so by appealing to the truth of materialism collapse under the difficulties of providing a formulation of materialism that does not hold it to be either untrue or unknowable. As such, the best that the developers of humanlike technologies can hope to achieve under the materialist paradigm is the creation of weak-AI: technologies that are humanlike insofar as they *emulate* human capacities. To claim that the development of strong-AI is possible under the auspices of the materialist paradigm is thus either an exercise of faith in the prevailing doxa, or it is the collateral effect of an endorsement of materialism made for reasons that sit beyond materialism *qua* materialism—i.e., the belief that endorsing a materialist metaphysic is a useful adjunct to the Promethean attitude.

In part two of the thesis, I challenge the idea that the technologies produced under the auspices of the materialist paradigm are capable of fulfilling the Promethean attitude's promise of a better quality of life. Instead, I argue the opposite—although such technologies *do* have their benefits in certain areas, these benefits come with the risk that we will find ourselves at odds with our future.

PART TWO:

THE SOCIAL UTILITY OF THE PROMETHEAN ATTITUDE, OR, WHY MATERIALISM'S TECHNOLOGIES WILL PROBABLY MAKE US MISERABLE

6 TECHNOLOGY'S ROLE IN MAKING US HEALTHIER AND WEALTHIER

Here we return to the evaluation of the Promethean attitude promised at the end of chapter 1. So far, we have seen that materialism most likely draws its assent from its proximity to the Promethean attitude—the idea that with the correct kinds of technology any problem is surmountable. This idea is intrinsically linked with metaphysical materialism and technological optimism. Materialism acts to support this attitude by projecting an image of the world as maximally manipulable, and this suggests that potential gains to our quality of life are without any conceptual limit: everything can be put to work in improving our lot. Technology is seen through an overly optimistic lens because it is the force through which we have—thus far—made our lives better on the whole. My claim that it is the Promethean attitude that acts to make materialism attractive (as opposed to materialism being attractive in and of itself) is supported by the fact that materialism struggles to offer a persuasive account of our subjective capacities. Consciousness, free will and selfhood are all obscured by the detached and neutral investigative methodology associated with material and functional analyses. In addition, the difficulties of finding a formulation of materialism that does not bind us to a metaphysic that is either false or unknowable compounds the unattractiveness of materialism. Taken in tandem, the methodological issues of materialism and the lack of a reasonable formulation of what it means to be a materialist support the idea that strong-AI—i.e., technologies in possession of a mind—are impossible under the materialist paradigm. It is a nonsense to argue that technologies made in the image of humanity that materialism most naturally espouses will be in possession of subjective faculties—they do not feature in the plans, so there is no reason to think that they would feature in the products.

The drive to create humanlike technologies—coaxed along by the Promethean attitude's promise of a better future—nevertheless still stands. In the remainder of the thesis we shall evaluate the potential impact that these technologies could have for our lives. If they promise to be of benefit, then perhaps the lack of a rational basis to an endorsement of materialism, in and of itself, can be overlooked on the grounds that the Promethean attitude that it helps to bolster is of continued social utility. If they instead risk a future of misery and anguish, we have good reason to be sceptical of the Promethean attitude. In which case, the dominant legitimiser of materialist assent falls. I begin in this chapter by looking at how these technologies can act to make us healthier and wealthier. In the next chapter, I shall look at how these technologies are poised to make warfare less ethical, bloodier, and more destructive. In the following chapter, I explore how the deployment of these technologies to the workplace and our governments threatens to strip us of our individual and collective agency.

6.1 PROMETHEAN PROJECTS I: HEALTHCARE

Perhaps one of the most likely candidates for demonstrating the social utility of humanlike technologies, such as weak-AI, is the impact they can have on acquiring and maintaining our health. This area is wide—ranging from economic efficiency in delivering healthcare (Topol, 2019., p. 49) to algorithmic analysis of protein interactions (ibid., p. 50)—so we shall narrow our focus on the area of potential impact most likely to be of consequence to the average person: finding and treating a medical condition.

6.2 THE POTENTIAL OF WEAK-AI IN MEDICAL DIAGNOSTICS

The process of diagnosing an ailment requires the acquisition and interpretation of data. In the context of general practice this usually occurs in the form of verbal reports of symptoms and physical examinations, but for more complex maladies a number of

diagnostic tests are used to aid the collection of information. Once in possession of the correct information, the healthcare provider will then, using their expertise and experience to guide them, interpret this to offer to the patient what they believe to be the most likely diagnosis. We can see that there are two key factors in giving an accurate diagnosis—a model of the patient, informed by the data collected about the patient, and the expertise and experience of the professional interpreting the model.

Though currently in the early stages of their development, a number of weak-AI systems are poised to be deployed in clinical settings to aid the diagnostic processes—already there is the suggestion that some of these systems have parity with medical experts in the diagnosis of diseases (Bjerring and Busch, 2020., p. 2). These successes have thus far been confined to diagnostics that rely upon clinical imaging—specifically, in the identification of skin and breast cancers from clinical images (Jiang et al., 2017., pp. 237 – 238)—but there is no indication that the diagnostic capabilities of these technologies are necessarily confined to the analysis of clinical images. Soon, it may be so that these diagnostic systems are able to compete with medical practitioners, not just in the analysis of "structured data [e.g.,] imaging and genetic data" (op cit., 2020., p. 5), but also in the application of "unstructured data" (ibid)—i.e., the knowledge gained from experience and engagement with the medical literature—to the diagnostic process.

As it stands, much of this unstructured data is not accessible to diagnostic weak-AI systems in the same way it is to a human practitioner: it requires translation from natural language into a format accessible to the weak-AI system, but progress in the development of algorithms to aid in this translation signals that soon this unstructured data could also be made available for diagnostic systems (ibid). Should this be the case, then weak-AI system

and practitioner parity could be replaced with weak-AI system supremacy. Medical practitioners require several years of training, and a lifelong commitment to professional development to remain apprised of developments in their field—in contrast, it has been suggested that already-existing weak-AI systems, given suitably translated data, are capable of processing the information contained within the medical literature at a pace that far outstrips human capacities.[24] With this in mind, it is possible to assume that diagnostic weak-AI systems "will eventually have at their disposal most of the medical evidence that is relevant in a specific context" (ibid., p. 6), thus expanding their capacities beyond those that any individual or team could wish to achieve.

6.3 THE BENEFITS AND THE DRAWBACKS

If a diagnostic weak-AI system has a grasp of much of the relevant medical literature and is able to use it to autonomously analyse the patient (or more accurately, a model of the patient) then the benefits of this are clear. Not only will a deep grounding in the medical literature help to minimise the risk of misdiagnosis, but such a diagnostic system would also likely be more reliable and time-efficient than a human performing the same role (ibid., p. 2). There will likely be other benefits too. For example, the system could be a useful tool for telemedicine (Chen, 2019., p. 255) enabling those who live in rural or remote areas with poor healthcare provision to access high-quality diagnostic services without the need to travel to see an experienced practitioner. The medical professionals themselves could also benefit from the deployment of diagnostic systems, becoming "less rushed, busy and hurried" (ibid., p. 254) as a consequence of the system shouldering some of the burden of their work. This could also free time to be used on patient care and rapport, and also lowers the likelihood of a stressed and overworked practitioner

[24] "A doctor reads about a half dozen medical research papers in a month … whereas [IBM] Watson can read a half million in about 15 seconds" (Captain, 2017., n.p.).

prescribing inappropriate medication (ibid., p. 253). Indeed, the system could also offer treatment suggestions itself based on recommendations found in the medical literature, thus lowering the potential for inappropriate or harmful prescriptions even further. It may even be so that the system could provide *bespoke* treatment recommendations, ensuring that each specific medication prescribed is maximally efficacious for that particular patient whilst also referring to their other prescriptions to avoid any potential conflicts—again, also reducing the burden on the practitioner who would have to consider this beforehand. All such benefits are of substantial desirability, and so it seems at this point, that the Promethean attitude's deployment of humanlike technologies to healthcare provision is justified. The application of technology to the problem of diagnosis is, conceptually at least, capable of bringing forward a situation where we can rely upon a system with capabilities that outstrip those of our expert diagnosticians, to provide us with reliable, efficient and easily accessible diagnostic services.

Yet things are perhaps not quite so clear-cut. Once deployed, these systems are likely to become integral to healthcare provision—their uptake spurred on by the benefits they promise to bring. But the widespread deployment of these diagnostic systems comes with risk attached. Perhaps the main risk that such systems would bring lies in the potential over-reliance on these systems. As the system would refer to the aggregate corpus of medical literature in its decision making, using the system would be something akin to consulting an expert—there would be a "*epistemic* obligation [for practitioners] to align their medical verdicts with those of the AI system" (op cit., 2020., p. 6). This deference of diagnoses to the system presents several difficulties.

Accuracy in diagnosis is important. Though the technological optimism of the Promethean attitude suggests that a weak-AI diagnostic system would be able to exceed the capacities of our very best medical practitioners, there are a number of factors that could impact upon the accuracy of the diagnoses that the system produces—especially so if the epistemic obligation is observed. The most obvious of these is that "poor quality data entry [leads] to unreliable data output" (Kilkenny and Robinson, 2018., p. 103)—i.e., the so-called "garbage in – garbage out" (ibid) problem. Inaccurate or incomplete data about the patient, either through human (e.g., the patient neglecting to report some important factor pertaining to their ailment) or machine error (e.g., an instrument with some unchecked manufacturing defect) impacts on the reliability of the data and this would naturally cause errors in the conclusions drawn by the system. Moreover, errors would also arise should the data used to train the weak-AI also be of poor quality—for instance, should inaccuracies be introduced into the unstructured data used to train the system during its translation into an AI-accessible format, or should such data be of inferior quality due to obsolescence or errors, then this would also cause inaccuracies to emerge in the system's diagnoses.

Relatedly, the information used to train the system may also contain within it some bias that then becomes "embedded in decision procedures" (op cit., 2019., p. 248). Should the training data draw from insufficiently diverse sources then clinically relevant factors—e.g., race, sex, etc.—could be neglected, thus leading to misdiagnosis. This is especially likely to occur in uncommon, difficult to diagnose conditions, whereby a lack of literature on the subject may inadvertently cause irrelevant characteristics to be given undue weighting in the system (Challen et al., 2019., p. 233). Say there is a condition that effects only a very small number of individuals, and all mentions of the disease in the literature happen to note

that the patient is male, then the system may claim that a female with an identical clinical presentation does not have the condition on the basis of her sex alone. In addition, it is also not possible to rule out the potential for interference for malicious actors motivated by xenophobia or political extremism deliberately introducing bias to the system in order to produce negative clinical outcomes for particular subsections of the population.[25]

Inaccuracy is not the only risk that diagnostic weak-AI systems threaten to expose us to; *too much* accuracy may also cause negative consequences. Training such systems with accuracy alone in mind means that they would have little appreciation for the "impact of false positive or false negative predictions within the clinical context of use" (ibid., p. 234). Compare, for instance, the identification and diagnosis of potentially cancerous skin lesions. When presented with an ambiguous case, because of relative ease of excision and the high risk of making a diagnostic error, a human medical practitioner is likely to err on the side of caution and recommend removal of the lesion: misidentifying a benign lesion as malignant carries fewer negative consequences than the inverse. Yet, this is unlikely to be so in the case of a weak-AI diagnosis; training the system to spot benign lesions without any consideration of the risk incurred with a false negative diagnosis means that malignant lesions may be misdiagnosed as benign. Such systems would need to consider the nuances of the diagnoses it produces, and the impact that these have on the patient.

6.4 THE BLACK-BOX DEALBREAKER?

The two above issues—that a diagnostic weak-AI may be inaccurate in some contexts, and *too* accurate in others—gestures towards the need for a human practitioner to supervise the workings of the system to mitigate the flaws that it may have. But the inscrutable nature of

[25] Recall the North Korean ransomware attack on the NHS in May 2017—such cyberattacks highlight the vulnerability of computer systems in healthcare services.

weak-AI systems makes this oversight difficult, if not impossible (ibid., p. 233 : op cit., 2020, pp. 4 – 10). Unlike simpler systems, which operate by passing an input through a complex, but epistemically accessible, set of 'if-then' rules, the operations of weak-AI systems are much more complex. Such systems take an input—e.g., information about the patient—and passes it through a "sequence of multiple hidden layers" (op cit., 2020., p. 7) that each manipulate the data. The outputted data from one layer is recursively taken as the input for the next layer, until the process reaches the terminal output layer. The precise manipulation of the data as it passes through the hidden layers—though informed by the data used to train the system—is epistemically unavailable to a human overseer. This so-called "*black-box* nature" (ibid., p. 3) of weak-AI systems is owed to their complexity—"the hidden layers in a network [can] contain millions of weights and thousands of distinct features" (ibid., p. 7), and this inscrutability is only compounded by extensiveness of the data used to train a system; one cannot expect even a group of senior practitioners to have a grasp on the entirety of the relevant medical literature.

There is therefore a tension here. As the system draws in more data from the literature, it theoretically becomes more accurate and robust, but the human oversight necessary to mitigate its flaws becomes more difficult as a consequence: a system claiming maximal accuracy will also likely be maximally inscrutable. This not only makes evaluating the accuracy of the system difficult—one would simply have to trial the system to ascertain its reliability—but it also means that "we risk losing medical understanding" (ibid). Put simply, because of the inscrutable nature of the technology, practitioners would be unable to understand how it reaches its diagnoses—some blind faith in the system's workings will likely be needed. But is this enough to undermine the usefulness of diagnostic weak-AI systems?

It does not seem so—the problem does not seem to be intrinsic to the system itself, but rather seems to be rooted in the practitioner's *relationship* with the system. The diagnostic work need not be delegated entirely to the system; a symbiotic approach could be taken instead. If the technology is used to *distribute* the burden of the diagnostic work, thus creating a cognitive-coupling across human-system boundaries (*à la* Clark and Chalmers, 1998) and—importantly—the practitioner treats this coupling as sacrosanct by remaining sceptical of the results it produces, then it does not seem that the inscrutability of the system presents an insurmountable barrier to its uptake.[26] If a practitioner denies the epistemic obligation that the system seems to place on them, and rather thinks of the system as a *tool* to aid in the diagnostic process—perhaps through suggesting differential diagnoses, recommending relevant treatments, etc.—then the inscrutability of the system becomes somewhat irrelevant. A practitioner does not need to know *how* an MRI machine, for example, works on a technical level to find it useful. Furthermore, confining the system to this supporting role means that the risk of losing medical understanding need not be present[27] and issues with inaccuracies and unnuanced diagnoses can be actively mitigated by the human practitioner whilst retaining many of the benefits that the system promises. As such, it seems that the Promethean attitude is, on balance, a valuable one to hold with respects to healthcare provision. Despite the problems that have been raised—not only can the deployment of humanlike technologies to healthcare improve patient access to high-quality diagnostic services whilst making the process less burdensome on practitioners, but

[26] Encouraging this scepticism perhaps may be the most challenging aspect of deploying these technologies, although it should be noted that the benefits to the working conditions of the practitioners may help to mitigate some of the circumstances—e.g., overworking—that are associated with relying on potentially damaging heuristics and dogma to inform diagnostic decisions.

[27] In fact, it is not an impossibility that the system could have the opposite impact—the system could deepen medical understanding by challenging the practitioner's own opinions and biases in much the same way as a colleague could.

it also can serve to reinforce medical understanding by challenging established dogma and heuristics, thus improving healthcare outcomes in the short and long term.

6.5 PROMETHEAN PROJECTS II: WEALTH AND FINANCE

The social utility of being in good health is clear and the role that humanlike technologies such as weak-AI can play in securing this utility is apparent. But health is not the only source of social utility—having one's finances in good health is similarly desirable (not at least because poverty impacts negatively on health). As we saw above, our humanlike technologies are especially suited to fields that rely on the acquisition and analysis of data—this suggests that management of finance is also an area which could benefit from the deployment of such technologies. As with healthcare, the applications of these technologies are broad, so we must narrow our focus—we shall examine the role that Promethean technologies can have regarding access to financial products and the growth of one's assets.

6.6 ACCESS TO FINANCIAL PRODUCTS: BARRIERS TO BORROWING

As noted in the literature "everything to do with understanding and controlling risk is up for grabs through the growth of AI-driven solutions" (Aziz and Dowling, 2019., p. 34). This has particular potential when it comes to access to financial products. Access to financial products is often contingent on ensuring that the recipient of the product represents a manageable risk to the vendor of the product. Consider an application for a bank loan. An applicant with healthy finances will be preferable to an applicant in financial difficulties, as the former represents less of a risk to the lender—there is less chance that the applicant will default on their payments and thus less chance that the vendor will lose their money. The financial health of an applicant, and thus the risk a vendor undertakes, is usually measured by way of their credit profile: an applicant already with good assets and a

record of repaying prior loans has easier access to financial products and often, these products are on better terms than those offered to someone with a less robust or extensive credit profile.

This arrangement presents a difficulty to those applicants that either have an insubstantial credit profile or lack one entirely. Lending to such applicants "is inherently risky as there is no history to draw on to check borrower reliability" (ibid): their access to financial products consequently suffers.[28] This situation, where financial products are inaccessible, entails a number of negative consequences, not only on an individual level, e.g., low levels of home ownership, an inability to weather temporary financial instability with credit, etc., but also on a societal level, insofar as development and consumption is hindered. Humanlike technologies such as weak-AI are capable of mitigating this barrier to financial products faced by those with an unsubstantial or non-existent credit profile. If the applicant grants the lender access to the information that they generate, that usually sits beyond the remit of a traditional credit check, then this can be used as input to evaluate the risk that lending poses (ibid., pp. 34 – 35). For example, the lender could request access to the applicant's internet browsing history: an applicant who frequently visits online gambling websites would likely represent a greater risk to a lender than another applicant who spends their time online checking local news. Other data sources could also be used to evaluate the risk associated with lending to the applicant—location data generated by the applicant's mobile phone could help to evidence regular employment, their transaction history could demonstrate the possession of stable income and online chat-logs could be

[28] In order to offer some estimate of the number of people that may be lacking in a sufficient credit profile consider that in 2017, 1.7 billion people globally remain without a bank account (Demirgüç-Kunt et al., 2018., p. 4)—disproportionately, poorer working women in the developing world (ibid)—and as a bank account is traditionally necessary for a profile to be built, this figure serves as a rough lower bound as to those effected. The actual figure is almost certainly higher, as it is likely that a subsection of account holders will not have a robust credit profile.

used to show that the applicant has a wide circle of friends and family who could act as informal guarantors to their loan. Indeed, because of weak-AI's ability to analyse large data sets, it is likely that, in practice, a vast range of these non-traditional data sources will be used to calculate the risk associated with lending (ibid).

Not only does this help applicants gain *access* to financial products should they lack an appropriate credit profile, but because of the automated and remote nature of these decisions, they are also less liable to be influenced by prejudice. In the traditional face-to-face image of lending, an applicant would need to have their application approved by a human representative of the lender and this presents a risk of biases and prejudices from representative—in regard to race, sex, sexuality, etc.—bearing upon the decision to lend to the applicant or not, and if so, on what terms. Whilst such discrimination is often legislated against, in practical terms, a dearth of alterative lenders—e.g., should one live in a so-called "financial-services desert" (Bartlett et al., 2019., p. 4)—and the costs associated with pursuing a mis-sold product through the courts, and so on, may expose borrowers to this kind of discrimination. Although, as we saw earlier algorithms are vulnerable to the incorporation of biases into their decision-making processes, the use of automated technologies to approve or reject the decision has *already* led to a more equitable arrangement for minorities accessing financial products—a reduction of roughly 40% in mortgage interest rates for minority borrowers is reported in instances where algorithms, rather than human actors, are in charge of the decision-making process (ibid., p. 16). One can assume that as these technologies become more embedded in the field, and the more equitable arrangements associated with them exert downward pressure on the price of financial products on the market—a borrower able to shop around will not choose a lender

that discriminates against them in their rates—then discrimination in access to financial products will be minimised.

From this, we can see that these technologies, when applied to this aspect of the financial world have clear social utility. Not only does it grant more people access to financial products which can then be used to personally or collectively beneficial ends, but it does so in a manner that also reduces discrimination and prejudice when accessing these products. Thereby these technologies are capable of resolving some of the structural inequalities that burden minority populations, allowing them opportunities to prosper that would perhaps otherwise not be available to them.

6.7 GROWING CAPITAL: TECHNOLOGY AND TRADING

Say one applies for a business loan: the loan is approved and on good terms. They start their business and accrue a reasonable profit. They want to expand their business so they can employ some of the graduates from the local university, but their profits will not cover the costs. Not wanting to apply for another loan, they reach out to a wealth management firm, who promises to invest their money—the returns on investment will help to increase their capital, bringing the day when they can expand forward.

The popular, traditional image of financial investment is one that is undoubtedly human—one can imagine the scene of a 1980s stockbroker on the floor of the exchange with a phone pressed to each ear shouting frenzied buy-and-sell orders to their colleagues. This image of investment has long been obsolete. Much financial activity has become automated: conservative estimates suggest “well over a half” (Wellman and Rajan, 2017., p. 616) of financial activity is automated, less conservative estimates suggest “all financial

markets and financial services are now automated" (Davis et al., 2013., p. 852). But what utility do humanlike financial technologies bring to the financial world?

To understand this, one first needs to understand the notion of arbitrage. Put simply, arbitrage refers to "a situation where a good can be sold in one market at a price higher than it can be bought in another" (op cit., 2017., p. 613). There are a number of kinds of arbitrage opportunities—for example, leveraging differences in exchange rates between currencies to turn a profit (ibid., pp. 614 – 615), or the buying and selling of shares for companies listed in more than one exchange—but to give us the best understanding of the efficacies of automated financial technologies, we shall focus on the more complex *statistical* arbitrage. Statistical arbitrage uses historical observation to identify "relationships among assets prices that hold probabilistically … but not by definition" (ibid., p. 616). Say for instance an electronics manufacturer releases a new product and it is well-received by reviewers and consumers alike; this translates to an increase in the asset price for that company. In this situation, it is not only the manufacturer itself that may experience a bump in asset prices—it is likely that those who make software or accessories for the product, or perhaps those who make the parts that go into the product will *also* see the benefit of the product's release. This represents an arbitrage opportunity. If the investor shorts the stock of the manufacturer— i.e., borrow stocks to sell immediately, only to rebuy the stocks to settle their debt once the market has calmed and they can be acquired at a lower price—and makes a long investment on the related stocks anticipating a rise in asset price, then the investor can turn a profit on this market behaviour.

As we can see, the successful leveraging of a statistical arbitrage opportunity depends on two factors. The first is discovery of the correlation in price between the two (or more)

assets. The second is the speed at which the buy and sell commands can be actioned. If the investor misses the window where the increase in asset price of the primary stock prompts an increase in asset price of the secondary stock, then the arbitrage opportunity vanishes. Autonomous financial technologies can outperform human agents with regard to these factors, thus helping to ensure the arbitrage opportunity is successfully leveraged (ibid). Machine learning, for instance, "expands the scope of data mining and data processing and thus, enhances the capacity to trawl markets in search of patterns and correlations to exploit" (Hansen, 2020., p. 3). By training the machine learning algorithm with real-life market data a model of the market can be produced, thereby enabling the discovery of correlations between price changes in a number of assets. This can not only be used to confirm intuitive relationships between assets—like in the example above—but can also be used to discover *unintuitive* relationships between seemingly disparate assets, thus identifying arbitrage opportunities that may have escaped the attention of a human investor. Furthermore, because the model is able to operate at "ultrafast" (ibid) speeds, it is able to make buy and sell orders in fractions of a second, thereby taking advantage of arbitrage opportunities that seem to appear and vanish simultaneously. Opportunities that would be impossible for a human agent to seize become possible for financial technologies.

6.8 DANGERS OF AUTONOMOUS FINANCIAL TECHNOLOGIES I: DUELLING AUTOMATED TRADERS

The ability of financial technologies to spot correlations between asset prices in the market and capitalise upon them with great speed, in addition to the relative affordability of computing hardware is part of the reason behind the shift to automated trading in recent years (op cit., 2017., p. 616). The other driver behind their uptake is their demonstrable utility in increasing profits for those that use them. So called "quant" (McAfee and

Brynjolfsson, 2017., p. 266) firms, i.e., those which use this kind of financial technology, have "built up spectacular records" (ibid) in providing high returns for those who invest with them—annualised returns for such firms have been reported to be as high as 12% (ibid., p. 267). A relatively low barrier to entry, combined with the incentive of higher returns has led to their widespread adoption—but this itself presents a risk. Due to the competitive nature of the field, the technologies employed are proprietary in nature. Keen to maintain an advantage over competitors, investment firms "are reluctant to share the recipe for the 'secret sauce' that gives them their competitive edge" (op cit., 2020., p. 3) and thus we see a proliferation of different market models and automated investment algorithms in competition with each other in slightly different, often overlapping, market contexts.

This means that competing models can interact with each other with unpredictable and potentially damaging effects. For instance, consider the so-called *Flash Crash* of May 6th, 2010, where in little under five minutes the equivalent of "around one trillion dollars" (Borch, 2016., p. 351) was lost from the US financial markets.[29] This sudden fall in the markets was precipitated by autonomous financial technologies—specifically, high frequency trading algorithms (HFTs), seeking to capitalise on fleeting arbitrage opportunities, rapidly buying and selling assets to and from each other (Sornette and von der Becke, 2011., p. 9). At the height of trading, 27000 of such transactions occurred in the space of fourteen seconds and this "translated [to] a negative spiralling effect" (ibid) on asset prices in the market-place—the HFTs had caught each other in a negative feedback loop. As we can see, this risk is one that comes with using such financial technologies: not

[29] Trading was temporarily suspended as a result, and once reopened, the market recovered almost as quickly as it declined (ibid).

only would a (rational) human trader not get caught up in such a loop, but even if they did, the speed at which they could action their trades would fall far below the nearly 2000 trades-per-second rate at which the HFTs operated, thus minimising the downward pressure on the markets.

The solution to this problem lies in ensuring that the market model that an automated trader uses is adequately adaptive to changes in market regime—i.e., changes in "the state or behaviour of a market at any given time" (op cit., 2020., p. 6). An accurate, sufficiently adaptive model will help to ensure accurate predictions and thus also helps to guard against the kind of negative effects seen above. Two difficulties are encountered during the development of an automated trading agent: underfitting, where bias is incorporated into the model so the automated trader "consistently learn[s] the same wrong thing" (Domingos, 2012., p. 81) and overfitting, where *noise* (i.e., random, irrelevant data) is incorporated into the model—the agent learns "random things irrespective of the real signal" (ibid) as a result. Both need to be overcome to ensure the reliability of the automated traders and to mitigate against negative effects, but, as was the case in a healthcare context, the opaqueness of such technologies makes it difficult to monitor the development of the model successfully. Recall that these systems have a large number of hidden layers that manipulate the input data in epistemically unavailable manners to reach their output.

In place of active human oversight, the developers of these technologies need to rely on heuristics to guide them—in a series of interviews with such developers, the principle of Ockham's razor was shown to be the guiding heuristic of choice (op cit., 2020., p. 7 – 9). In this context, the application of Ockham's razor manifests itself in a preference for the

simplest possible model for the task at hand: "you always want to default to the model with the least complexity" (ibid., p. 7). This aversion to unnecessary complexity helps to minimise the risk of under- and overfitting by ensuring that the models stay as comprehensible (ibid) and intuitive (ibid., pp. 8 – 9) as possible. The more complex a model is, the more difficult it is to understand it and the more difficult it is to control it and mitigate against any errant data incorporated in the model. Simpler models are easier to understand and so present a lower risk when deployed: human actors do not have to contend with overwhelming opaqueness to bring them back under control. Whilst the risk of competing autonomous agents interacting badly with each other may never be totally eradicated, keeping models simple can go some way towards minimising the chances of negative impact—in the light of the power of these technologies to grow one's wealth and the social utility of a wealthy populace, this is perhaps a risk that we should be willing to live with.

6.9 DANGERS OF AUTONOMOUS FINANCIAL TECHNOLOGIES II: DISTANCE

The drive to maintain an edge over the competition does not just risk multiple autonomous trading agents interacting badly—it also presents a problem regarding individual or organisational responsibility and the ethical regulation of financial markets. We have seen, there is an incentive for quant firms to keep their technologies secret and this, in combination with the potentially opaque nature of these technologies and the fact that "finance is generally not understood as a technological (and social) practice" (Coeckelbergh, 2015., p. 287) leads to a problem with "distancing" (ibid). To understand what is meant by distancing in this context, it is perhaps best to briefly revisit the Heideggerian concept of *aletheia* explored in §2.9. Recall that *aletheia* refers to the process of "bringing-forth … out of concealment into unconcealment" (Heidegger, 1954., p. 317). Distancing is much the inverse of this. Instead of bringing-forth the financial

world into unconcealment—bringing something obscure into clear relief—there is the danger that the continued uptake of autonomous financial technologies will obscure that which was once (relatively) clear. As financial technologies progress and become more complex, operating with greater autonomy it will become progressively more difficult to ascertain precisely what they do, and *why* they do it.

This is an issue as it violates the long-held criteria for responsible action. In order to ascribe responsibility to an action it must not occur "by force or through ignorance" (Aristotle, *NE*: III.i, 1110a)—i.e., "you have to know what you are doing and you have to have control over what you are doing" (op cit., 2015., p. 288)—and thus it is not immediately obvious where the responsibility for the actions of autonomous financial technologies would lie. Clearly it cannot be the technologies themselves, for they operate not knowing what they do, and neither can they choose *not* to do something, e.g., on the basis of some ethical principle, because ultimately, their actions are the brute product of complex mathematical wrangling. Similarly, the developers and users of the technologies seem to escape responsibility too. By using these technologies "the people who make trading decisions and exchange goods remain out of sight and can therefore escape responsibility ascription" (ibid)—because these technologies are oftentimes autonomous and cryptic, the risk is that those that use them are neither in direct control nor in possession of complete understanding of the technologies.

This difficulty in ascribing responsibility to the acts of autonomous financial technologies leads to difficulties in regulating the financial markets. Not only does their distancing effect make it difficult to ascertain what needs to be done—i.e., how to change the technologies to ensure they conform to ethical standards—or who should bear the

responsibility if something untoward were to happen, but it also makes the *democratic* oversight of financial forces more difficult too (ibid). If financial technologies reach a level of complexity where their workings evade the understanding even of those with expertise in the field, then the prospect of the electorate-at-large understanding the technologies—a vital component of reasonable democratic oversight—trends towards impossibility.

But can this issue with responsibility and oversight be overcome? The issue has two aspects to it: one that focusses on knowledge, and the other, control. Using Ockham's razor as a heuristic for the development of autonomous financial technologies is also valuable when tackling the knowledge aspect of the problem. By always preferring the simplest model, the potential for understanding the logic of the technologies is maximised. Although the inner-workings of the technologies may resist total understanding, maximising comprehensibility in the development and deployment of these technologies goes some way towards mitigating the distancing effects that the technologies precipitate. But what of the problem regarding control? *Prima facie* it seems that the solution we found for the problems of humanlike medical technologies—i.e., the introduction of a "judgement call [before] … the final decision" (Svetlova, 2012., p. 430)—seems to mitigate the lack of control human agents may have over the technology whilst simultaneously giving us a way to ascribe responsibility for its actions—whomever makes the call becomes responsible. Yet this solution undermines one of the major benefits of deploying these financial technologies in the first place: speed. If the automated agent must wait for a human actor to sign off its actions, then the ability to capitalise on a fleeting arbitrage opportunity may be lost and thus this solution is not particularly attractive.

To settle the issue regarding control whilst retaining the benefits of the technologies, the role of the human must shift from the authoriser of automated actions to the *organiser* of automated actions (op cit., 2020., p. 10). As the organiser of the automated actions, the human is charged with the task of deciding what information is relevant and thus should be included when training the model, and in addition, they must also interpret the quantitative data that the system outputs. The technologies can identify patterns, but it cannot give *meaning* to these patterns nor understand the relationship (if any) that gives rise to the patterns: "you need to have a human, who understands how the relationships work" (ibid). For example, the technologies may spot a spurious relationship between two different factors—maybe the price of a particular asset seems to fluctuate in line with the phases of the moon—the user of the technology must guide it so it does not act on these spurious relationships. By undertaking this guidance, the human actor does not just exercise some, albeit indirect, control over the technology, but also allows us to ascribe responsibility to them—if something goes wrong (or right), it is owed to the quality of the human actor's ability to organise the technology.

In this chapter we have seen how these technologies are poised to make us healthier—by giving us better access to high quality diagnostic services that simultaneously lower the burden on healthcare providers—and wealthier—by making financial products easier to access and allowing us to increase returns on our investments through the exploitation of arbitrage opportunities at machine speed. To continue the evaluation of the Promethean attitude I will next explore the impact that such technologies may have on the battlefield, to demonstrate how these technologies—though capable of making us healthier and wealthier, are also capable of making our lot worse in other ways, and—arguably—worse overall.

7 AUTOMATED WEAPONS: THE PROMETHEAN ATTITUDE AT WAR

So far, we have seen that humanlike technologies, developed to do the things that we can do, are capable of not just improving our health, but also improving our wealth. Yet social utility does not lie solely in material improvements on our condition—maintaining a stable and secure international order is also a source of social utility. Being at war is detrimental to all involved, even those most tangentially included. In addition to the burden borne by those actively engaged in combat, there is also a political and economic cost to waging war that must be borne by the citizens of the nations, or organisations, involved. Due to the high-stakes nature of war, and the magnitude of suffering that it entails, warfare presents itself as problematic enough to attract the attention of the Promethean attitude. This attention has resulted in many advances in the technologies used to wage war—indeed, one may trace the advent of the bow and arrow and similar obsolete weapons to the same Promethean drive—and if the social utility of such endeavours is to be demonstrated, it must be so that the technological products of the Promethean attitude act to minimise the burden that warfare necessarily brings.

7.1 THE KILLER ROBOTS WE HAVE AND THE KILLER ROBOTS WE WANT

Currently, autonomous weaponry—for example, close-in weapons systems, deployed as point-defence and tasked with the interception of incoming missile fire—is commonplace in the arsenals of developed militaries. But the autonomous weaponry deployed today does not have *true* autonomy; most, if not all, currently deployed autonomous weapons systems (AWS herein) are either "human 'in the loop'" (Scharre and Horowitz, 2015., p. 8) systems—whereby the weapon "use[s] autonomy to engage … targets that a human has decided are to be engaged" (ibid), such as guided munitions, etc.—or are "human 'on the loop'" (ibid) systems, whereby the weapon autonomously selects and engages its own

targets but with “human controllers [that] can monitor the weapon system’s performance and intervene to halt its operation if necessary” (ibid), as is the case with close-in weapons systems. The goal of the AWS developers is to create a system that can run without any meaningful human oversight. A truly autonomous AWS would decide which targets to attack, and it would do so in a way in which “human controllers cannot monitor… performance [or] intervene to halt [the AWS’] operation if necessary” (ibid). With an AWS that is truly autonomous, human oversight and direction would simply be redundant.

This prompts the question of *why* we would want to create weapons systems with no human direction or oversight —there seems to be no *prima facie* reason for wanting this, yet there does seem to be a *prima facie* reason *against* their creation. Put simply, losing one’s life at the behest of a lethal algorithm deprives the deceased of a certain sense of dignity. Caught in the sensors of an AWS, a combatant is reduced to the cold input of an equation. Their significance becomes expressed purely in mathematical terms—a variable to be acted upon. As such, the use of AWS seems to violate the ethical maxim intuitively possessed by those who are hostile to the idea of AWS and articulated neatly by Thomas Nagel: “whatever one does to another person intentionally must be aimed at him as a subject, with the intention that he receive it as a subject … it should manifest an attitude to *him* rather than just to the situation” (Nagel, 1972., p. 136). It is hard to see how a person, dehumanised via a process of being reduced to a mathematical input, could be recognised as a subject in such circumstances.

The violation of this principle manifests itself (though there are also likely other contributing factors) in widespread hostility towards the very notion of creating AWS. For example, a 2013 opinion poll asking the US public whether they opposed the creation and

use of these weapons revealed that the idea of AWS is held in low regard across all demographics and political identities (Carpenter, 2013., n.p.).[30] If the proponent of AWS is to make their case, the ethical benefits and the social utility of AWS must be of such a magnitude as to make this dehumanisation a bearable cost to their deployment: we must gain more from their use than we lose.

7.2 THE PUTATIVE ETHICAL BENEFITS OF AUTOMATED WEAPONS

Proponents of AWS point to its potential ethical benefits in order to argue for their deployment. In their view, AWS are well positioned to bring a number of advantages over human combatants that will lower the burden associated with the prosecution of war. Of these proponents, Ronald C. Arkin is perhaps the most optimistic about the benefits of AWS. For Arkin, the benefits of AWS fall into two broad but interlinked categories: those that are rooted in the positive abilities of AWS, in and of themselves, and those that come from the fact that AWS is resistant to many of the weaknesses associated with human combatants. We shall explore these in turn.

Arkin notes that robots are "already, faster stronger and … smarter than humans" (Arkin, 2010, p. 333) and that these improvements on human capacities means that they are preferable to human soldiers for the purposes of warfare. In his view, AWS will be able to better achieve the goals of war and they will do so whilst "treat[ing] us more humanely on the battlefield than we do each other" (ibid). By way of example, consider a situation in an asymmetric war where a soldier seeks to engage a poorly identified target: it *looks* like

[30] Perhaps the most interesting statistic revealed by the polling is that nearly two-thirds (65%) of those currently serving in the US military (at the time of the poll), and half (50%) of those who had previously served reported strong opposition to AWS. When including all reports of opposition by current or ex-military personnel (those strongly opposed and those somewhat opposed), the proportion of respondents opposed rises to nearly three-quarters (73%) for those actively serving and nearly two-thirds (63%) of those with previous military service (Carpenter, 2013., n.p.).

they are a legitimate target, but they cannot be sure from where they stand. The soldier has three options: they can engage the target, thus running the risk of killing a civilian; they can keep a watchful eye on the suspected target; or, they can change position to get a clearer view of the target, but at the risk of being spotted and engaged by an enemy lookout. This is not the case for a mobile AWS. Because of the expendable nature of AWS, the risk of death in the third option above is avoided—there is no loss of life if the machine is destroyed—so there are no prohibitive risks in seeking a better vantage point from which to identify the enemy: "there is no need for a 'shoot first, ask-questions later' approach" (ibid). Not only does this guard the soldier against unnecessary risks but it also has, in Arkin's view, the potential to decrease the chances of accidentally targeting and engaging a civilian.

The ability to physically get closer to a potential target is not the only benefit that AWS could bring to the battlefield. Arkin also suggests that the accuracy of such systems can outstrip the accuracy of human soldiers by virtue of the fact that they can be equipped with a "broad range of robotic sensors better equipped for battlefield observations than humans currently possess" (ibid). For instance, an AWS can be equipped to sense radar signatures, detect infra-red thermal radiation or determine the presence of footsteps in an adjacent room through delicately calibrated seismic sensors. These sensors do not need to be part of the unit itself; remote sensors can be consulted with much the same speed and fluidity as those hard-wired into the unit itself. This ability to consult a variety of different sensors, self-integrated or remote, that gives the AWS abilities unable to be possessed by human soldiers is also amplified by the ability to "integrate more information from more sources far faster before responding with lethal force than a human possibly could in real time" (ibid). Information received from sensors could be quickly checked against intelligence

collected about combatants to minimise the chance of civilian casualties, and moreover, there is nothing to prohibit several AWS forming a network, sharing the flow of battlefield information between them without the lag associated with human-to-human communication. These supra-human capacities for acquiring and integrating battlefield data with other sources of data, in concert with the acceptable loss of the AWS in combat gestures towards a future whereby AWS can "perform more ethically than human soldiers are capable of performing" (ibid., p. 334), with far fewer instances of collateral damage and a less-dangerous battlefield for friendly human soldiers.

Supra-human abilities are not the only attractive features of AWS according to Arkin—part of their draw lies in the fact that they are specifically not human. Unlike humans, AWS are not vulnerable to the types of human failure that cause the most egregious kinds of war crimes (ibid., pp. 334 – 338). Examples of these failings are plentiful: religious and ethnic hatred, binary *good-vs-evil* thinking, and fervent nationalism all precipitate profoundly immoral crimes, such as genocide, torture, and wartime sexual violence. An AWS would not be motivated to commit such atrocities—the AWS would be impervious to hatred. Similarly, an AWS would not seek vengeance—a factor Arkin notes is "related to an increase in ethical violations" (ibid., p. 335)—nor would it be susceptible to fear, panic, anger, or any other emotional stresses that could increase the incidence of ethically unsatisfactory behaviours: "robots … would be unaffected by the emotions, adrenaline, and stress that cause soldiers to overreact or deliberately overstep the Rules of Engagement and commit atrocities" (Lin et al., 2008., p. 1).

7.3 TWO CAVEATS TO TEMPER ARKIN'S OPTIMISM

Free from the psychological burden of battle, the AWS would operate in a reliable fashion, and equipped with the supra-human capacities for sensing their environment and quickly

analysing diverse streams of data, the AWS seems to present a marked ethical improvement over a traditional human soldier, yet this is not without caveats. Although Arkin believes the advantages of AWS are owed to what it is intrinsically, and thus the deployment of AWS are necessarily beneficial, the potential costs incurred if he is mistaken are high. If he is wrong, unintended loss of life is likely. Umbrello et al (2020) seek to temper Arkin's optimism by suggesting that two conditions must be met before AWS should be deployed.

Firstly, it is imperative that only "moral" (Umbrello et al., 2020., p. 276) AWS should be used in warfare, whilst "non-moral [AWS] should be prohibited" (ibid., p. 277). In their view, this notion of a moral AWS does not mean that the AWS is developed in line with one of the established schools of ethical thought[31] as there is too much disharmony between the schools to "[establish] a universal theory of morality that would be accepted by all lawmakers" (ibid., p. 278). Instead, in their view, moral AWS are those developed to adhere to the already-existing *legal* frameworks governing conduct in warfare (ibid., p. 279). These frameworks are twofold: firstly, the internationally observed *jus in bello* laws, which provide the international legal baseline for conduct in conflict, and secondly, the domestic national laws that set out the *Rules of Engagement* governing wartime conduct for that particular nation, which may contain more strenuous conditions for wartime conduct than the international *jus in bello*. In their view, using already-existing legal frameworks to guide the development of AWS, not only avoids a time-consuming and challenging process of building an international mandate for a unified code of ethics, but

[31] Although there are those that would disagree with this and argue that existing schools of ethical thought *should* guide the development of such technologies. For example, Vallor (2016) offers an interesting argument for developing technologies (including autonomous weapons) in line with a virtue ethicist's framework.

also guards against many of the most extreme potential consequences of deploying AWS, e.g., using them to commit genocide.

The second caveat recommended by Umbrello et al., is that the deployment of AWS should be delayed until the technology is able to reliably discriminate between a combatant and a civilian in the area of operations (ibid., p. 275). Not only do civilians enjoy legal protections from engagement, as stipulated by *jus in bello,* that demand reliable discriminatory capacities in order to be observed (an AWS developed to respect internationally agreed conduct but lacks the ability to actually do so is problematic) but accidentally engaging civilians because of poor discriminatory functionality undermines much of the utility that AWS promise to bring: an unreliable AWS would make warfare much more dangerous.

7.4 TWO THEORETICAL PROBLEMS OF AWS: ABSURD WAR AND MAXIMALLY STRESSFUL WAR

Let us grant that the caveats set by Umbrello et al are met, and AWS are deployed to the battlefield. We consequently find that they perform better on the battlefield than even the staunchest optimist may have hoped. In such a scenario, there is no *operational* advantage to deploying human combatants to the battlefield as they perform worse than the AWS, and there *is* an ethical incentive *not* to have humans on the battlefield—i.e., to protect them from harm. Now let us assume that the war is fought between two well-developed militaries, both of which make heavy use of AWS. The operational and ethical incentives hold for both sides. Human soldiers are no longer in active combat as there are no good strategic reasons to have them on the battlefield, and there *are* good ethical reasons against it. Such a war would instantiate the vision of the "kind-hearted people" (Clausewitz, 1832., p. 13) mentioned by Clausewitz who "think there was some ingenious way to … defeat an

enemy without too much bloodshed, and might imagine this to be the true goal of the art of war" (ibid)—a maximally ethical war with no loss of life on either side. But this kind-hearted vision of a zero-casualty war impedes the ability to achieve success in war whilst also leaving the impetus for prosecuting war intact.

If we borrow Clausewitz's understanding of the purpose of war—"*an act of force to compel our enemy to do our will*" (ibid)—we can see how this is so. Two nations fighting a just war (according to *jus ad bellum*) will have exhausted the non-violent means available to them to achieve this goal. Diplomacy, economic sanctions, etc., will have all been pursued in vain: war is the last resort. The drive to impose one's will over another remains, but the use of AWS in the place of human combatants means that the ability to impose of one's will on another is impaired. Recall Arkin's argument that AWS present an advantage over a human soldier because they are more expendable than a human soldier—the loss of an AWS is preferable to the loss of a human life. It is this expendability that impedes the imposition of will: one cannot make another do what one wishes by destroying things that they do not mind losing. This problem from expendability is exacerbated by the "relatively low production cost" (op cit., 2020., p. 274) of AWS. As they are cheaper to produce than conventional weaponry and equipment[32], the argument that prosecuting war would be economically debilitating cannot be easily made. In this scenario, where war is fought between AWS alone, the war would be reduced to a display of power in the hopes that the adversary is dissuaded from prosecuting war, but if this grandstanding is not successful, as the expendability of AWS suggests, then the automated battles would simply be the absurd opening spectacle before either a traditional, human-vs-human, or, in the case where one

[32] Many of the desirable abilities of AWS comes from the software that they run "which [can] be copied practically for free" (Scharre, 2018., p. 130)—naturally, this means that much of the cost of AWS is incurred by its hardware, which would benefit from an economy of scale if standardised and mass produced for an army.

army's AWS have been destroyed (alongside their capacity to manufacture more) an asymmetric AWS-vs-human, conflict takes place.

This absurd situation is perhaps unlikely. Let us look at a more probable scenario. Let us argue that the performance of AWS, though better than humans in certain regards, is not so good as to render human combatants redundant. In such a situation, wars are fought by hybrid teams—human personnel are supported by AWS, and this is also the case for one's opponents. A conflict fought in such circumstances, because of the efficacies of AWS, will likely be much more dangerous for human combatants than a standard, human-vs-human war. Not only would AWS be better at targeting and neutralising combatants, but the AWS' lack of emotion also risks increasing the danger for combatants. With the capacity to be motivated by hatred, attrition or fear, also goes the capacity for mercy, clemency and kindness. All targets identified will be eliminated—the AWS, unable to show mercy, clemency or kindness is unable to change their course of action: they will kill the target if they can. This gross increase in lethality, in the context of fighting an unfeeling machine, will do nothing to mitigate the very same mental stresses that the AWS would, in part, be deployed to relieve. Indeed, it is imaginable that fighting against AWS, does not only *not* contribute to lowering stresses, but may also bring about a *more* stressful situation than soldiers would normally face. It seems sensible to suggest that the lack of interpersonal connection (as highlighted by Nagel above) and overall higher lethality would imbue the human soldiers more readily with the panic that leads to irrational and ethically undesirable behaviours. In which case, the supposed ethical gains of deploying AWS could be offset by the ethical losses caused by increasingly fraught human soldiers, thus bringing the utility of AWS—at least with regard to the supposed ethical advantages—into question.

7.5 A PRACTICAL PROBLEM: ROBOTS AND INTERPRETATION

As Umbrello et al., suggest the deployment of truly autonomous AWS requires, if it is to be of any serious utility—ethical or otherwise—the ability to accurately select its own targets. In the context of war, targets can be of two basic kinds: personnel, i.e., soldiers, and military objects[33], i.e., "those objects which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage" (Additional Protocol I, 1977: Article 52). Therefore, a useful AWS would need to be able to reliably identify such targets in the complex milieu of a battlefield. Moreover, in order to "satisfy the *jus in bello* requirement of proportionality, autonomous weapons will [also] need to be able to identify and enumerate *civilian* targets reliably" (Sparrow, 2016., p. 98; emphasis mine). This is because, according to the requirement of proportionality, engaging a target likely to cause significant collateral damage (e.g., shelling a residential area of a city because of intelligence that an enemy commander is directing operations from somewhere within) is forbidden, and naturally, the number of civilian deaths likely to be caused by engaging a target cannot be estimated if the AWS cannot accurately identify civilians. From this, we can see that a useful AWS must be able to identify three kinds of entity—enemy soldiers, military objects and civilians.

The identification of such entities is complicated somewhat by the fact that not all enemy soldiers are *legitimate* targets. *Jus in bello* stipulates a number of circumstances where an enemy soldier —even if correctly identified—is not able to be legally engaged. One such circumstance we have already seen: the engaging of an enemy combatant in a situation

[33] There are those that suggest the legal definition of *military objects* also extends to human beings (e.g., Dinstein, 2007., pp. 84 – 85) but for the sake of the argument here, I have separated human combatants out from inanimate objects that possess military advantage as to better capture the complex task facing AWS.

likely to cause considerable collateral damage is forbidden (ibid., p. 99). An enemy combatant is also an illegitimate target if attacking them would "constitute an unnecessarily destructive and excessive use or force" (ibid). For example, if a soldier shot *through* a civilian to incapacitate a sniper behind them then this would constitute a violation of this law. Furthermore, an enemy combatant becomes an illegitimate target if they are *hors de combat* (ibid)—an injured, unconscious, or unarmed enemy combatant is classed as *hors de combat* and thus is an illegitimate target, as is also an enemy combatant who expresses their wish to surrender. In short, an enemy combatant is not a legitimate target simply because they are an enemy combatant. There is a further layer of abstraction in identifying a legitimate target, insofar as there are further criteria that must be met in order to legally engage the target.

Similar complications are also found in the targeting and engagement of military objects—the inanimate entities that can confer military advantages to those involved in war. This is again because the legitimate targeting of a military object is not owed solely to what the target *is*. The permissibility of engaging the target relies not only on its "nature, location, purpose or use" (Dinstein, 2007., p. 85) but also on whether "its destruction, capture or neutralisation … offer[s] a definite military advantage" (ibid). To explain this, let us borrow an example from Heather M. Roff—that of a church. As Roff notes, a church is "normally given protected status" (Roff, 2014., p. 216). Under normal circumstances, a church is not a permissible military target as its destruction would violate *jus in bello*. But, say if one of the belligerent parties used the church as a warehouse to store their ordinance then the change in the use of the building, in combination with the clear military advantage of destroying the enemy's ordinance, changes the status of the target and thus makes engagement permissible (ibid). As such, as is the case with the targeting of military

personnel, there are further criteria which must be met before targeting a military object becomes legitimate.

7.6 HEIDEGGER AND TARGETING I: THE ONTIC-ONTOLOGICAL NATURE OF TARGET IDENTIFICATION

From this we can see that the AWS needs to perform two tasks in order to ascertain accurate, legally permissible targets: first, it must work out what *there is* in the battlefield (i.e., it must be able to reliably tell the difference between a soldier and a civilian, a barracks and a church, etc.), and secondly, it must work out the *significance* of what there *is* in the battlefield (i.e., whether a soldier is unconscious or surrendering, or whether the destruction of a military object will confer them any advantage, etc). We can see that there is some significant resonance here between the task of the AWS identifying legitimate targets and the task of Heidegger's Dasein grappling with the meaning of being. Recall the example of the heir sorting through their inheritance in §2.9. To make sense of the objects bequeathed to them, the heir first engages in an ontological enquiry. They work out what is before them by grouping discrete elements of the milieu before them together on the basis of the "fundamental structures" (Heidegger, 1927., p. 8) revealed to them by the objects themselves. Once this ontological enquiry is completed and the heir knows what is before them, the heir moves on to discover ontic knowledge about the objects; they begin to grasp the significance of the objects in relation to a particular field of understanding, and in doing so the "particular domains of knowledge are exposed and delimited" (ibid). This is precisely the task that befalls the AWS when autonomously selecting and engaging its targets; it must first do ontology to work out what there is, and then it must grasp the ontic significance of what there is, in order to establish whether or not it is a permissible target to engage. Without being able to do both of these interpretive tasks, the AWS risks

misidentifying its targets, and thus acting in a way that undermines the supposed ethical advantages that such technologies are purported to bring.

7.7 HEIDEGGER AND TARGETING II: CAN ROBOTS DO ONTOLOGY?

Heidegger's framework for performing an ontological investigation—i.e., the revealing of the *what-being* of an object, via distinguishing its fundamental structures from the fundamental structures of other co-present entities—is not altogether dissimilar from the method by which an AWS (or any other humanlike technology that can 'see' what is before it) attempts to recognise the objects interacting with its optical sensors. The primary difference pertains to the nature of the significant fundamental structures—they are different between a human and a machine. Whereas a human may be concerned with the form of an object, its fragility, weight in the hand, and so on, the machine instead uses *mathematically* significant features of the image produced by its sensors. The machine may look for instances in a digital image where "there is a 'significant' change in image brightness' (Russell and Norvig, 2010., p. 936), using the differences in the mathematical values for each of the pixels to identify the edges of the object in the image, or the machine could "[break] an image into regions of similar pixels … associated with certain visual properties, such as brightness, color and texture" (ibid., p. 941) in order to detect the boundaries of such regions, and thus the shapes of the entities in the image.[34] These techniques are used to identify significant features—the fundamental structures—of the image. This structural information is then, to simplify the process, checked against a bank of other images with similar structural information, but which have been previously

[34] These are two of the more basic techniques used by machines to 'see'—the techniques employed in the most up-to-date technologies are considerably more complex, though are still based on the same process of using the mathematical values that make up a digital image to identify relevant features of the image.

correctly identified (usually by humans[35]) to note the degree of similarity between them. If the image in question shares a high degree of mathematical similarity with the known image (or parts thereof) then the machine can be confident that the two images are of the same thing. As such, it *seems* that an AWS performing the task of isolating the relevant structural information and checking it against known images is, at some level at least, capable of the kind of ontological investigation that Heidegger talks of.

This said, there are two issues that prevent us from saying that humanlike technologies can do ontology, and which also might make the notion of AWS deeply unattractive. The first is that the ability of a machine to do ontology—that is, work out what there is in its environment—seems to be inversely proportional to the complexity of the image which it is trying to identify. Although simple images can be identified with a relatively high degree of accuracy, the more complex the situation becomes, the more difficult it is to reliably work out what there is: "it is notoriously difficult for a computer to reliably identify objects of interest in a given environment … [and this] is even more the case in crowded and complex unstructured environments and when the environment and the sensor are in motion relative to each other" (op cit., 2016., p. 98). This presents an issue for AWS, as not only will they be deployed in the complex, unstructured and dynamic environment of a battlefield, but they will also need to be self-mobile to meet their supposed utility. Because of this, it is likely that the AWS will—at best—be unable to identify little in its environment with any desirable level of accuracy, or—at worst—*misidentify* entities on the battlefield, thus running the risk of ethically undesirable behaviours (e.g., the accidental

[35] Oftentimes, this dataset is crowd-sourced. For example, whenever one has to solve a captcha (e.g., by selecting all the images that contain, for instance, streetlights or motorbikes, etc.) to prove that one is a human in order to submit information to a website, one is helping to build the dataset of known images. It is telling that the identification of these objects is proof enough that one is not a robot.

killing of civilians). It should be noted though, that this is perhaps not an argument for abandoning the prospect of AWS altogether but is instead one for delaying their deployment—it may be so that one day, the AWS's capacities for accurate discrimination of environmental entities may reach human parity, and if this happens, this point becomes moot.

The second problem is more detrimental to the project of developing and deploying AWS to the battlefield. There exist certain kinds of images—adversarial examples—that can be created by researchers "that exploit deep neural networks' vulnerabilities to trick them into confidently identifying false images" (Scharre, 2018., p. 180). These images can either be (what appears to a human as) meaningless static, which is then confidently identified as being an object by the technology, or images that appear to humans to be of one thing—for example, a dog—and to a weak-AI, to be another—for example, a bunch of flowers. Adversarial examples do not have to be cleverly manipulated digital images. Researchers have also created 3D objects that are similarly capable of tricking weak-AI image identification algorithms (Athalye et al., 2018, p. 284)—for instance, researchers created a 3D printed model of a sea turtle that was repeatedly misidentified by weak-AI as a rifle (ibid). Adversarial examples present a particularly important challenge to AWS because they are notoriously difficult to guard against. Even if an adversarial example is fed back into the system so it can learn to ignore it in the future, this only goes as far as to protect the system from that *one specific example*: "because the space of all possible images is 'virtually infinite' … many more fooling images can be evolved … [thus] no matter how many fooling images the AI learns to ignore, more can be created" (op cit., 2018., p. 185).

To guard against adversarial images effectively, researchers would need to understand the workings of the technology but, as we have seen elsewhere (§6.4 and §7.9), comprehension of such is hindered dramatically by the opaqueness of the system's internal logic (ibid., p. 182). Therefore, adversarial examples present a difficult, possibly insurmountable challenge for all technologies that require accurate image identification but is of special significance to AWS because of their lethality—mistakes in target identification readily translates into unlicensed loss of life. It is not impossible to envision a situation in a battle where one side could exploit adversarial examples to their advantage. These 'spoofing' attacks could be defensive—insofar as one side could use them to mask their own strategic interests, using adversarial examples as a form of AWS-specific camouflage—or more seriously, they could be offensive. It is not hard to imagine a desperate or unscrupulous actor, under the cover of the plausible deniability conferred to them by the opaqueness of the humanlike technologies, marking prohibited targets with such adversarial examples in order to provoke their opponents' AWS into committing atrocities—indeed, in light of the example of three-dimensional adversarial objects above, doing so seems to require little more than giving a child a suitably manipulated toy sea turtle to play with.

The ethical implications here are clearly undesirable—because of the vulnerable and lethal nature of AWS, adversarial examples present themselves as a largely undetectable way to mark a target for elimination, regardless of the legal or ethical permissibility of that target. It is also difficult to protect against such instances without also including a human (with capacities for ontological investigations that outstrip those of AWS) somewhere 'in the loop'. But deferring the ontology to humans, which effectively this would do, undermines many of the supposed benefits of AWS—e.g., full autonomy, ability to operate at machine

speed, etc.—and thus is an unattractive solution to the problem of adversarial examples. The capacity to fall foul of spoofing attacks is therefore good enough reason to find the deployment of AWS unattractive, but for the sake of argument, let us assume the faith in technology that characterises the Promethean attitude. Let us suggest that at some future date, with some future technology, the problems regarding ontological investigations are resolved—AWS have reached a stage where their ontological enquiries produce results with parity to humans.

7.8 HEIDEGGER AND TARGETING III: CAN ROBOTS GRASP THE ONTIC SIGNIFICANCE OF OBJECTS?

As Heidegger suggests, when grappling with the question of being, grasping the ontological nature of an object—its *what-being*—alone does not give a proper appreciation of what the object *is*. This is because ontology forms the foundation of ontic significance. Grasping the ontological does not mean that we automatically grasp the ontic: one might be able to recognise a bust of Beethoven but this does not mean that they know that he was a composer. Being able to grasp the ontic significance of an entity is of crucial importance for AWS, because the *legitimacy* of a target is an ontic understanding of the object. There is nothing *given*[36] to us by the *what-being* of a church or a soldier that renders it a permissible target or not. If this was the case, the (im)permissibility of engaging the target would be an enduring condition of that target—an individual soldier, for example, would not be able to flit between being a legitimate or illegitimate target if such was a feature of their ontology. It is therefore necessary for AWS to grasp the ontic significance of an entity if it is to guard against any illegal or unethical mistargeting—but can an AWS do this?

[36] As we saw from the discussion of Sellars (1997) and Quine (1951) in §2.6, it is impossible for the legitimacy of the target to be *given*, for there are no such things as givens—all understandings are a product of interpretation.

In order to assess this, we must first recall that Heidegger holds Care to be the fundamental source from which our ontic (and ontological) understandings are drawn from. It is through Care that we interpret our worlds and the beings within them and, importantly, this includes ourselves. By interpreting the world, through either mode of Care, we seek to answer the question of our own being—humanity "takes a stand on itself" (Dreyfus, 1991., p. 43) by giving meaning to that which it finds itself amongst. This question of self-interpretation—the concern with one's own being—is the wellspring from which Care comes forth: without this concern with one's *self*, it is difficult to see how Care and the interpretation of entities could arise. It is precisely the importance of the self to our interpretive endeavours that prohibits the AWS from being able to properly grasp the ontic significance—and thus the permissibility of engaging—its targets.

Perhaps the best way of illustrating this point is to contrast the AWS with a human soldier, in the context of interpreting the permissibility of targets on the battlefield. A human soldier is conscious, and from this consciousness they also find themselves in possession of a self and of freewill. These three facets of the human condition intermesh with each other: from their consciousness emerges both free will and their selfhood, which then becomes an issue for them. From their free will emerges the capacity for Care which affords them the opportunity for world- and self-interpretation which in turn, allows them to address their concern with their own being. As we saw in §2.7, addressing the issue of their own being rests on the ability to erect their own network of significance, the form of which is influenced by the projects that they set for themselves. One of these goals, in this instance, may be to be a good soldier, and if so, the soldier becomes motivated to grapple with the ethical considerations surrounding combat. Gaining familiarity with the legislation and the ethics of war is a requisite for interpreting oneself as a good soldier.

But then how does the soldier realise this theory in practice—with the theoretical distinction between permissible and impermissible targets in place, how does the soldier then interpret the world in such a way to make it fall into line with the theory? This depends on what the target in question is. If the target is a human, then the soldier can circumspect the Care the enemy has for their world to recognise them as another conscious being *doing* something. Say for instance, we are concerned with the discrimination of a surrendering enemy combatant—a human being will recognise this by first way of recognising the "sameness of being" (Heidegger, 1927., p. 115) between themselves and their opponent, specifically, with regards to "what [their opponent] does, needs, expects, has charge of [and] in the things at hand which [they] initially *takes care of* in the surrounding world" (ibid., p. 116). This recognition of sameness, revealed by how the Other engages with the world reveals to the soldier their intentions—in short, the soldier does not recognise a surrendering soldier through, for instance, the brute presence of a white flag, but rather by recognising that displaying a white flag is a function of Care, through which their opponent expresses their intentions. At the point of spotting the flag, the soldier can recognise that there is an Other *being* present, for which the flag is meaningful and thus the intended meaning of the flag, as revealed by the Care of the Other, can be recognised. The soldier can do something similar with regards to potential militarily significant structures such as churches—the soldier can recognise that the structure occupies a location in an Other's network of significance by observing the way in which they are involved with it. In this way, they can also recognise when this meaning changes in a way to make it a legitimate target: filling a church with ordinance demonstrates to the soldier that their opponents have ceased to see it as a site of spiritual significance.

An AWS is unlikely to have the capability to do this for two reasons. Firstly, recall that the image of humanity that guides the development of humanlike technologies is impoverished: it cannot properly grasp human consciousness (chapter 2), free will (chapter 3) and selfhood (chapter 4). As such, developing maximally humanlike technologies—strong-AIs—is impossible when operating in a materialist paradigm. The best we can do is to create technologies that *emulate* human capacities—i.e., weak-AIs. But this means that the humanlike technologies we *can* produce, are those that do *not* possess consciousness, free will and selfhood—precisely the human faculties from which our ability to Care originates. As such, an AWS *qua* weak-AI, is ill-equipped to perform ontic investigations. The ability to interpret a target as permissible or not, is contingent on faculties that the AWS simply does not possess. Proponents of the Promethean attitude will likely argue that this is unimportant, and that there are ways around the issue—perhaps by weighting certain variables manually or designing the technology with certain ethical goals in mind. This leads us to our second reason.

It is important to remember that our humanlike technologies operate primarily through mathematics, and if the above argument is to be undermined, the AWS—or rather, its developers—must find some way of creating a mathematical analogue to the human capacity for Care that also does justice to the malleability and fluidity associated with Care. It is not enough to code hard rules for the AWS to follow, for they must be able to adapt to changing contexts to be of any serious use in a battlefield. Yet translating Care into mathematics—creating a Care algorithm—seems to be an impossibly difficult task. Care is a more fundamental phenomenon than the creation of algorithms: it is *through* Care that we create algorithms—creating algorithms and humanlike technologies is a way of being involved with the world; it is a way of *Caring about* things.

Thus, the question becomes one of whether it is possible for a product of Care to get in touch with the Care that produced it in order to *emulate* it. This seems unlikely. Although the products of Care bear the hallmarks of the Care that produced it[37] they are necessarily *subordinate* to, and thus *distinct* from the Care that produced them. An algorithm is *secondary* to the Care that produced it and has no way of transcending itself in order to reach the same ontic-ontological status as the human: human handiwork does not, and cannot, share the same way of being as the human that produced it, by virtue of the fact that it lacks consciousness, free will and selfhood. Any endeavours to translate consciousness, free will and selfhood into algorithmic form will be derailed by the inability to adequately capture qualitative experience in a quantitative form—it is qualitative experience that informs the evaluation of a target's legitimacy. Until the AWS can adequately grasp the qualitative—something that its very nature currently prevents it from doing—it is not possible to say that an AWS can perform the kind of ontic inquiry necessary for accurate target discrimination. All they can do is attempt to *emulate* our capacities for ontic inquiry, but due to the dynamism of what it is that they are emulating, an emulation will never be as useful as the *first-hand* ability to Care about our worlds.

In light of this inability to grasp the ontic, the issues surrounding adversarial examples interfering with its ontological capabilities, and Nagel's aforementioned objection on the grounds of a dearth of respect, the deployment of AWS constitutes a risk of unethical behaviour that is of such a magnitude as to offset the supposed ethical gains from their deployment considerably. As such, we see here that the development of such automated weapons is of poor social utility—put simply, the Promethean attitude and its humanlike

[37] Such informs Maurice Merleau-Ponty's concept of "cultural objects" (Merleau-Ponty, 1945., p. 348ff).

technologies, in a context where subtle nuances and interpretive accuracy and flexibility are key, can make an already distressing and dangerous situation even more so for those involved.

In the next chapter, I continue the argument that these technologies are poised, ultimately, to make us miserable by exploring their impact on the worlds of work and governance: the gains in health and wealth that such technologies can bring are offset not only by their making warfare less ethical, bloodier and more distressing, but also their ability to truncate and mar our individual and collective agency.

8 THE IRONY OF THE PROMETHEAN ATTITUDE

We have so far seen that when applied to discrete fields, the Promethean attitude's utilisation of humanlike technologies—those that *emulate* human capacities, such as weak-AI—can, in some contexts, be of positive social utility. In others, they are of negative social utility. We have yet to explore the effects of the Promethean attitude, *at large*. In order to understand how the *broad-spectrum* impact of the Promethean deployment of humanlike technologies will ultimately foist us into a situation of relative misery, we shall explore the *irony* of the Promethean attitude—that the drive to control our world via humanlike technologies ultimately undermines itself. We shall then examine how this operates in two key fields—the world of work, and the world of politics—in order to demonstrate that our humanlike technologies have the potential to undermine human agency, which then calls their social utility into question.

8.1 THE IRONY: ONE STEP FORWARD, TWO STEPS BACK

Let us refresh our understanding of the Promethean attitude. Gains in the quality of life bought about by technological progress leads to an optimistic disposition towards technology. This sense of optimism then finds itself bolstered by materialism, as it serves to extend the scope of technological optimism to all worldly phenomena. As everything is somehow owed to the material, either by virtue of its being material or through its supervenience on the material, and matter can be manipulated by technology, there is nothing that cannot be utilised or improved by technology. In other words, the Promethean attitude manifests itself in an "unlimited confidence in the ability of humans and their technologies to overcome any problems" (Dryzek, 2013., p. 52). From this, we can see that the idea of control is central to the Promethean attitude—the world is reified as a resource that we can bring under our dominion to achieve our goals. This attitude has been the

driving force behind much of our technological advancement. Recent humanlike technologies, e.g., weak-AI, machine learning, etc., have been developed with the aim of maximising control over the world. Through spotting patterns in data and then acting on such patterns, these technologies allow us—in theory—to expand our control over the world. Yet it is here—in this notion of control—that we begin to see the *irony* of the Promethean attitude: as we make our technologies ever more humanlike in their capacities, we risk losing the very control that the Promethean attitude seeks to expand and our agency becomes diminished as a result.

Before we see how this manifests itself in practice, it is important to note why—in theoretical terms—the Promethean attitude is vulnerable to this irony. The fact that the Promethean attitude is, at its heart, the synthesis of metaphysical materialism with technological optimism means that its products are well placed to precipitate agency undermining effects. Recall in §3.8, where it was shown that a commitment to materialism encourages a neglect of the existential embeddedness of agency, thus forcing materialist accounts of free will into defending a form of free will characterised either by reactivity and pettiness, or illusion. This inadequacy in accounting for free will is then transposed into the Promethean attitude. Though the producers of humanlike technologies perhaps do not spend much time thinking about metaphysics, when they do, materialism—by virtue of its hegemonic nature (as we saw in chapter 1), and the ease with which it is integrated with their existing Promethean mindset—is standing by to offer its authority. But this passive commitment to materialism means that the adherents to the Promethean attitude are at best ill-equipped to deal with, or at worst, completely ignorant of, matters regarding human agency.

This alone is sufficient to bring the attitude's abilities to properly safeguard agency into question, but these doubts are also compounded by another aspect of the attitude—technological optimism. The overly optimistic image of technology possessed by the Promethean attitude—i.e., that of an overwhelming social good (Brayford, 2020a., p. 527)—leads the adherents of the attitude to become blind to technology's negative ramifications (ibid., p. 529). Thus, the Promethean attitude engenders an unusual situation: a passive commitment to materialism means that adherents are improperly equipped to deal with matters of agency, but at the same time, their optimistic view of technology interferes with their ability to reckon with the issue—if technology is a necessary good (at least in the long term), then how can it have a negative impact on our agency? As such, matters of agency do not naturally show up for the Promethean attitude, but even in the case that they did, the attitude's passive commitments to materialism prevent it from dealing with such matters appropriately. The materialist's image of agency is distorted to the extent that, even if forced to confront it, the producers of humanlike technologies could not do justice to the concept. Therefore, insofar as this blind spot exists and is obscured behind technological optimism any Promethean endeavour—i.e., any efforts to use technology to bring the natural world into our dominion—is capable of surreptitiously undermining human agency.

8.2 UNDERMINING AGENCY I: AGENCY AND MEANINGFUL WORK

To see how these technologies undermine agency, we should clarify our understanding of the term. Borrowing from chapter 3, the account of agency employed here is not just the brute capacity of being able to choose one course of action or another (although this is certainly part of it), but rather is one that also pays homage to the existential significance of our choices. Agency is more than just choice: it is choice that is bound up in reaffirming one's consciousness—"our selfhood and our enduring identity" (Tallis, 2020., p. 191).

Agency refers to the ability of an individual, embedded in a particular context, to make decisions that bear existential weight: agential actions are those that contribute to the sense of *who* a person is, and what meaning they find in their existence. This is not to say that the possession of agency infers that a person is able to act in any way that they please—agency does not suggest an absence of constraint (ibid., p. 218)—or that for one's actions to count as free, they must be explicitly and consciously executed in their totality (ibid).[38] Rather, what really matters is that the individual bears ultimate responsibility for whatever decisions they choose out of those afforded to them *and* that this responsibility is somehow integrated into *who they are.* The capacity to possess a robust sense of identity or the ability to add meaning into one's own life is contingent on the possession of agency, because, without agency we have no ownership over what we do—if our actions are not *ours* then they cannot form part of who we are. With this understanding in place, we can turn our attention to *how* technology can bring about a situation where individual agency is undermined.

Agency is intrinsically linked to the development of an identity and our ability to draw meaning from our lives. As a considerable portion of our lives is spent at work, it makes sense that work—provided it engages properly with our agency—"can contribute to a meaningful and flourishing life" (Veltman, 2016., p. 105). How can work do this though—what are the conditions that must be met before our work carries suitable existential weight? Susan Wolf, in her work regarding meaning in life, suggests that work becomes enriching—a source of meaningful self-understanding—when the nature of the work is such that the value of the work is "at least partly independent of my own existence and

[38] As Tallis points out, a choice to go to the pub is not rendered 'unfree' because we do not "consciously … contract all the muscle fibres involved in walking" (Tallis, 2020., p. 218).

point of view" (Wolf, 2012., p. 43).[39] This is to say that, for the work to be seen as meaningful—i.e., existentially enriching—it is not enough that the worker has *chosen* the work because it aligns with their self-image, it also must be linked with some kind of extrinsic value.[40] The subjective value of the work must be entwined with the *objective* value of the work if it is to avoid the dearth of meaning associated with "totally egocentric" (ibid., p.41) activities, i.e., those "devoted solely toward the subject's own survival and welfare" (ibid). Meaningful work must not *just* align with one's own wishes. It must also gesture to something of objective value, something that lies beyond the confines of the subject. Wolf does not go far into elucidating what would constitute such objective value: her suggestion is that it has something to do with being "useful" (ibid., p.36) and offering an opportunity for an individual to "develop her powers, or realize her potential" (ibid). We must turn to Veltman to add further detail into the idea of meaningful work.

Veltman argues for the existence of "four primary dimensions of meaningful work" (op cit., 2016., p. 117). The first of these borrows from Wolf's suggestion that meaningful work is related to the development of an individual's abilities—if work "develop[s] and exercis[es] the worker's human capabilities" (ibid) then the work is more likely to be seen as objectively meaningful than work that does nothing to advance a worker towards self-actualisation. As per the second dimension, meaningful work is likely to bolster feelings of "self-respect, honor, integrity, dignity, or pride" (ibid), and it also, as per the third dimension, is in some manner expressive of a "personal purpose, or … a genuinely useful purpose for others" (ibid). Finally, the fourth dimension of meaningful work is found in the

[39] Wolf does not specifically talk about work—her focus is on our activities in general, but naturally, this includes our work.

[40] 'Chosen' is used in a broad sense here—the choice need not be active or explicit: responsibility can be ascribed to the worker by way of tacit assent.

"integrat[ion of work with] elements of a worker's life, such as by building or reflecting personal relationships and values or connecting a worker to an environmental or relational context with which she deeply identifies" (ibid). Therefore, according to Veltman, work can be classed as meaningful if it is both concordant with subjective meaning—i.e., the work has to be *subjectively* valuable for the worker—and is in dialogue with the four dimensions of objective value noted above.

Veltman's dimensions of meaningful work reveals to us that meaningful work is primarily concerned with worker self-actualisation and the subsequent creation of an enduring identity, both in the public sphere—insofar as the worker is recognised for their achievements and the contributions that their work makes for others—and in a private sphere—insofar as the worker draws a sense of self-respect, pride, etc., from their work. It is clear how this resonates with our understanding of agency as the cultivation of a meaningful, maximally actualised identity is the *raison d'être* of agency: it is by choosing to take on the responsibilities associated with one's work that one becomes able to develop one's skills, engender a sense of self-respect, engage with purposive *good works* and find the expression of one's own relationships and values in what they do. In short, meaningful work is replete with the existential weight that is integral to matters of agency. Yet, this process of self-actualisation and identity formation is complicated by the introduction of humanlike technologies to the workplace. We shall look at two such complications: one that pertains to the gross unavailability of work in a technology-saturated labour market, and another that examines how our technologies can render work meaningless. We shall explore these in turn.

8.3 ARISTOTLE'S SHUTTLE AND THE GLOBAL WORKFORCE

The idea that we can create technologies that perform tasks to such a standard that they render the labour of humans unnecessary is an old one—the earliest mention of the idea is perhaps found in Aristotle's *Politics*. Here, Aristotle speaks of an imaginary situation "in which each instrument could do its own work…a shuttle would then weave of itself, and a plectrum would do its own harp-playing … [thus] managers would not need subordinates and masters would not need slaves" (*Politics*, I.iv, 1253b23). Aristotle's imaginary world of autonomous shuttles and plectra began to become a reality in the late 18th century—as we saw in §1.5, Adam Smith notes that human labour can be "facilitated and abridged by the application of proper machinery" (Smith, 1776., p. 13) which in turn maximises the productive capacities of a workforce. Equipped with the right kinds of technologies, the productive capacities of human workers can be increased by several orders of magnitude, thus creating ever more efficient productive procedures and turning ever larger profits as a consequence. This provided the impetus for the adoption of labour-displacing technologies in the workforce—the more machines and the fewer workers an employer employs, the greater the profits that a firm could create.

The name of this worker displacement by machines—technological unemployment—was coined by John Maynard Keynes. The link between the development of our technological capacities and the displacement of human labour was clear in his view. He agrees with Smith's observation that technology offers us the "means of economising the use of labour" (Keynes, 1930., p. 80), but importantly he also notes that this economisation "outrun[s] the pace at which we can find new uses for labour" (ibid). Put simply, the rate at which human labour is rendered redundant is higher than the rate in which new jobs are created. Once technology has negated the need for their labour, the displaced worker finds it difficult to find employment elsewhere. The development of advanced technologies—

weak-AI, ICTs, Machine Learning, etc.—has expanded the threat of technological unemployment. Whereas the technological unemployment seen in Keynes' day and earlier mostly displaced jobs in the manufacturing sectors, leaving many roles in other sectors unscathed, the adaptability of contemporary humanlike technologies means that many more sectors are vulnerable to technological unemployment.

Although jobs that involve "explicit and quantifiable rules" (Kim and Scheller-Wolf, 2019., p. 321) are, *prima facie* "more easily codifiable [and] hence more likely to be automated than work that involves … unquantifiable knowledge" (ibid), as the capacities of our technologies grow "the distinction between more easily and less easily codifiable world is quickly blurring" (ibid). Advanced algorithms, robotics and weak-AIs are poised to be able to perform tasks that we have traditionally believed to be too "cognitive" (ibid) for non-human actors to perform.[41] As such, large sections of the labour market are vulnerable to automation—when asked about the vulnerability of 70 different job roles, representative of the entire spectrum of employment in the USA, AI and robotics experts predicted that roughly 40-45% of the roles are at risk from automation (Walsh, 2018., p. 639). This is largely in agreement with other sources that report similar estimates of vulnerability to automation (op cit., 2019., p. 321): as our technologies improve the number of jobs vulnerable to automation is likely to increase.

It is not only the expansion of automation and the subsequent technological unemployment that exerts pressure on the workforce—technology has also facilitated globalisation. Technological advances, from the mundane—e.g., standardised shipping containers—to

[41] We have already seen two such examples of this—recall the discussion in chapter 6 about weak-AI in healthcare and in finance. The analytic work performed by the technologies resists easy codification and is too voluminous to be undertaken by human workers, but the unique nature of these technologies makes such work possible.

the advanced—e.g., mass communications technologies—have enabled ever greater economic integration between countries: “technological progress reduces transportation and information costs—thus promoting globalization” (Guriev, 2018., p. 201). Whilst this has caused a considerable improvement in living conditions and a reduction in rates of poverty in some of the poorest countries across the globe, globalisation exerts pressure on the workforces of developed nations. This is because technology has allowed employers to take advantage of workers in jurisdictions with less rigorous worker protections than those in developed nations—the payroll costs incurred by employers are lower and, as a result, profit margins are increased. This effect is especially pronounced in “electronic offshoring” (Ford, 2015., p. 118)—where “jobs are moved to low-wage locations instantly and at minimal cost” (ibid).

Again, humanlike technologies threaten to intensify this process by bolstering the competitiveness of lower-skilled offshored workforces in relation to higher-skilled domestic workforces. With access to the correct kind of technologies, lower-skilled offshore workforces can augment their analytic capacities in order to make higher-skilled work accessible: “a smart young offshore worker wielding such tools might soon be competitive with far more experienced professionals in developed countries who command very high salaries” (ibid., p. 121). This means that, in instances in jobs where total automation is either difficult (e.g., in complex legal cases rooted in abstract concepts, such as libel or privacy), undesirable (e.g., as is the case with medical diagnoses, as seen in §6.4), or where human oversight is otherwise necessary, employers can utilise low-skilled and offshore employees to perform this work—thereby lowering costs and increasing profits.

8.4 THE SHIFT IN WORK: NO WORK, BAD WORK, GOOD WORK?

These two phenomena—technological unemployment and offshoring—are posed to have a profound effect on the nature of our work, which in turn could have a similarly profound impact on our ability to choose meaningful work that contributes to our sense of identity. Perhaps most obviously this operates by way of its effects on the brute availability of work. Both technological unemployment and offshoring contribute to "job polarisation" (Goos et al., 2014., p. 2509)—that is a reduction of availability of jobs for mid-skilled workers. The displaced workers then either find work in low-skilled sectors (where work either resists automation due to its nature, or already entails low-enough renumeration to make automation or offshoring unattractive), 'skill-up' to join the high-skilled workforce or join the economic surplus population. As such, the chance that an individual finds themselves in meaningful work decreases in line with gross job availability—as profit is the primary impetus behind automation and offshoring, there seems to be no good grounds to assume only unedifying jobs will be lost. Meaningful work will likely be lost at the same rate as non-meaningful work.

This though, is not the only way that technology threatens access to meaningful work—technology also threatens to undermine existing meaningful work by removing responsibilities from human workers and shifting them onto technologies. As technology becomes increasingly embedded in the workplace, the proportion of the labour burden that it shoulders also increases. If we follow this trend to its logical conclusion, a situation emerges where all existent work is either resistant to automation and offshoring—because of its nature, or because it is economically inefficient to do so—or consists almost entirely in verifying the work done by weak-AIs, algorithms, etc. It is hard to see how this latter situation—where the once diverse and multiple responsibilities taken up by workers are, in

effect, reduced to oversight—can be conducive to the notion of meaningful work as, arguably, it violates every dimension of meaningful work as articulated above by Veltman.

The first dimension—the development of human capacities—is severely impeded if it is weak-AI that takes up most of the workplace responsibilities, as the opportunity for a worker to dedicate "time and practice" (Veltman, 2016., p. 119) towards honing their skills, or developing their capacities would be non-existent—the technology, not the worker, is that which does the work and so it is the capabilities of the machine-learning weak-AI that are developed, not the worker's. A similar point also holds true for the second dimension: if it is the technology that does much of the work, rather than the worker, then it is not clear how a worker could gain any sense of self-respect, honour, integrity, dignity, or pride, through their work as these would only belong to the worker if they were responsible for whatever act engenders them. The idea that one may derive a sense of pride from their proximity to a machine doing what it has been designed to do is a nonsense—when a student passes their maths exam with high marks, their resultant pride is owed to their command and understanding of the appropriate mathematical operations and formulae, not because their calculator did not malfunction that day.

How automation affects the third dimension—purposiveness—is less clear. Although one could argue that the worker could derive some meaning from their diminished overseer role in a by-proxy fashion—that is if the work they oversee "genuinely contributes something of value to the lives of others" (ibid., p. 125)—but this is difficult to square off with the notion that "being needed" (ibid) is a necessary factor in "contributing to a broader totality beyond oneself" (ibid). For example, consider the possibility that the weak-AIs performing the work have capacities that outstrip human workers—they perform their

tasks with an accuracy beyond that of even established experts in the field—and so human oversight is perhaps only required to sate some manner of morally, legally or politically mandated principle. In this case, there is no *practical* need for the overseer and this is enough to violate the purposiveness of the overseer's role as their role is effectively reduced to box-ticking. Finally, the fourth dimension—that which pertains to meaningful work's ability to "render an individual's life coherent … [by] integrating elements of his life … [and] integrating a worker in an environmental or social-relational context with which he identifies" (ibid., p. 132)—is also undermined by technological change. By negating the need for much human labour, technology diminishes the chances that a worker will find themselves in any context where social or environmental integration can occur. Oversight of technology does not require an extensive workforce and the profit-seeking impetus behind the deployment of technology to the workplace means that inefficient (i.e., overstaffed) teams are unlikely to be common. This drive for efficiency limits the opportunities that a worker has to ingratiate themselves in a particular social context: fewer co-workers entails fewer opportunities for social bonds to form. Similarly, in the context of the drive towards efficiency and profit, it is hard to envision this kind of overseeing work occurring in such an environment that allows the worker to "[sustain] relationships with loved … places" (ibid., p. 134) as ICTs allow this oversight to happen from any location with sufficient connectivity, thus decontextualising work from its environment. This is perhaps most keenly felt in the case that an offshore worker is responsible for such oversight: it is not impossible that an offshore worker is charged with the oversight of work that is situated in a very different context from their own. For example, a worker in a firm in New Delhi may oversee work produced by weak-AIs housed in a data processing plant in rural Pennsylvania, in which case, the actual location

of the work is so far disconnected from the worker to defeat any attempts to derive or strengthen any meaning pertaining to geography.

This image of future employment, characterised by a lack of jobs, and an even greater lack of meaningful work, is perhaps too pessimistic for some. There are those that would charge such an image as being an example of the "luddite fallacy" (op cit., 2019., p. 321). Belief in the luddite fallacy—i.e., the view that technology necessarily causes a decline in the gross availability of work—is seen by critics as arising from ignorance of economic history. They would point towards the fact that the "dynamic nature of capitalism has always leveraged technology to create more jobs than those that were lost" (ibid). Historically, deployment of technology has spurred a gross *increase* in job availability, not a decrease, and more jobs increase the likelihood of finding meaningful work through the brute force of numbers. This line of thinking is problematic though, for it does not properly grasp that our contemporary humanlike technologies are dissimilar in nature to those that have existed historically. In the past, when a worker is displaced by technology, the shift into one of the newly created roles usually necessitates some kind of reskilling—as new roles open, workers need to acquire the skills needed to perform said role. Naturally, this process of reskilling takes time. Workers in the past were often granted the time needed to reskill and find new work—the technologies stood in reserve until an adequately skilled workforce was in place. Yet, this is likely not going to be the case in the future. Because "humanity has recently become much better at building machine that can figure things out on their own" (McAfee and Brynjolfsson, 2016., p. 139), this process of reskilling is now a race against machines—if a weak-AI can be trained for a newly created job more quickly, or at lower cost than a human can, then it will not matter how many new jobs are created as human workers will still find themselves crowded out of the workforce.

From this we can see that humanlike technologies, when deployed to the workplace under the auspices of profit maximisation, are likely to have a negative effect on the availability of meaningful work. Whether this operates by way of reducing the gross availability of work or by reducing the responsibilities of a worker simply to oversight, the end product is invariably a situation where meaningful work is *less likely* to be found. This interferes with the formation of an individual's sense of self by way of reducing their agency: their ability to choose work that contributes to their identity and the meaning they find in their life. Humanlike technology's presence in the workplace either closes off an individual's potentialities of action by dictating that they are to take up the role of unfulfilling oversight, or alternatively, it denies them access to a situation where they can act in such a way that is congruous with the four dimensions of meaningful work—one cannot find *meaningful* employment if one cannot find employment *at all*.

This reduction in agency is made more pronounced when one also factors in the fact that the situation is one that has befallen workers. The workers themselves have not chosen to deploy the technologies that constrain their ability to engage in meaningful work: it is something that has happened *to* them. Technologies have been deployed to workplaces in order to maximise profits, but this effectively forces workers to cede their activities over to the technologies and their agency is left impoverished as a result—no longer are they able to engage in work that hones their skills, from which they can draw a sense of pride, which helps them to contribute to a purpose larger than themselves or reaffirms their relationships to people and places that are already meaningful to them. As such, this situation is a manifestation of the irony spoken of above: making our technologies more humanlike—e.g., by making them capable of performing 'human' labour—comes at the cost of losing

some of our agency, which in turn has a detrimental impact on the formation of identity and meaning that relies upon that agency.

8.5 UNDERMINING AGENCY II: DEMOCRATIC ACCOUNTABILITY

We have seen that humanlike technologies, when applied to a workplace context, are able to undermine our agency by lowering the availability of meaningful work. As such this process, in effect, undermines *individual* agency (or at least one important aspect of such). In addition to this, humanlike technologies are also able to undermine the *collective* agency vital to the proper functioning of our democracies. The drive towards efficiency (i.e., in the context of business, profit) and a sense of technological optimism lent to us by its effects in raising our standards of living (as explored in chapter 1) has led to humanlike technologies becoming embedded in business—the same also holds true for our governments. Attracted by the efficacies seen in a business context, our governments have also begun to embed these technologies in their operations: "both governments and companies use data-driven algorithms for decision making and optimization" (Lepri, et al. 2017., p. 11). These algorithms can be deployed to any domain in which relevant data can be or has been collected. For example, data can be pulled from historical crime records—e.g., the location and type of crime, the background of the perpetrator, etc.—combined with "aggregated and anonymized mobile phone data" (ibid., p. 7) and analysed by an algorithm to discover "the existence of multi-scale complex relationships in space and time" (ibid) which helps law enforcement to predict the likelihood of future crimes occurring in specific locations and in specific time frames. This information can then be used to better allocate resources across a region—say for instance, the data reveals that very few crimes have been reported in a specific location, in a specific timeframe, but another has frequent reports of crime over the same period. In such a case, it makes sense to redeploy resources focussed on the former area to the latter to combat crime with greater efficiency and efficacy. Similarly, resources

deployed at times with low incidences of crime, can be redeployed at those with high incidences of crime, thus making more efficient use of already existing resources. Similar applications can be found elsewhere in the provision of public services, from deciding where best to allocate government investment (ibid., pp. 8 – 9) to traffic management in city centres.[42]

Prima facie, the use of algorithms in government decision-making procedures seems to be the kind of overwhelming good that technological optimists espouse: greater efficiency means better public service provision. This though is not necessarily the case. The delegation of decision-making to technology, in the interests of greater efficacy, can undermine our collective agency and our ability to hold our governments democratically accountable. To understand how this arises, we must first introduce the notion of the "*responsibility gap"* (Matthias, 2004., p. 176) that is present when algorithms and other Promethean technologies are introduced into decision-making processes. As we saw in §6.9, the standard Aristotelean view of responsibility holds that an actor is responsible for their actions under the condition that they both know what it is they are doing, and they have control over what it is they are doing: "Things that happen by force or through ignorance are thought to be involuntary" (*NE*. III.1110a). Yet, the deployment of humanlike technologies to decision-making roles interferes with our ability to ascribe responsibility to an actor as the automatic nature of such technologies, in concert with their

[42] Perhaps the most prominent currently existing example of this kind of data-driven governance is found in the People's Republic of China. Their Social Credit System takes personal data, e.g., "ID Numbers, employment records, and education data" (Liang et al., 2018., p. 419) and combines it with data gleaned through surveillance methods—e.g., browsing data, financial data, etc.—to create a comprehensive digital record of an individual's life. This record is then leveraged by algorithms to make decisions regarding a number of public provisions. Citizens or organisations with a poor social credit score may find themselves paying higher repayment rates for loans, having business permits revoked, etc., whereas those with good social credit scores can have priority access to public services, pay lower business fees, etc. (ibid., p. 433).

ability to develop their own capacities "inevitability lead[s] to a partial loss of … control over the [technology]" (op cit., 2004., p. 176).

For instance, let us say that an AWS, of the kind seen in chapter 7, takes the life of a civilian. In such an instance, any actor to which we would try to ascribe responsibility will be able to deny responsibility with some substantial degree of plausibility. The operator of the weapon can deny responsibility on the grounds that they are the operator of the weapon in name only: the choice to engage the civilian was due to the weapon's own software. Similarly, if we are then to try and hold the programmer of the weapon responsible, they can escape this by highlighting the fact that the technology is adaptive and learns to select its own targets and so the decisions taken by the weapon are owed to factors that exist beyond their involvement with the weapon.[43] There is nobody with sufficient control over the technology to be able to ascribe responsibility to them. This problem is also compounded by the opaque nature of humanlike technologies—as we saw in §6.4, §6.9 and §7.7—the complexity of these technology impacts negatively on their comprehensibility, further widening the responsibility gap. As the "clear and distinct symbols" (ibid., p. 181) found in traditional software are, in these technologies, replaced with a "matrix of synaptic weights, which cannot be interpreted directly" (ibid), even if there was a way that we could say that one particular actor was in control of the technology, they would be able to escape responsibility by appealing to their ignorance of the precise intricacies of the technology's inner workings.

When humanlike technologies are introduced into the decision-making processes of governments, this difficulty in ascribing responsibility carries over into a political context.

[43] These two examples are based on two of the examples found in Matthias (2004., pp. 176 – 177).

In a standard representative democracy, electors in a particular district elect a representative, who then exercises their judgement on legislative matters on their behalf: power that resides in the electorate at large is lent to their representative. As such, ascribing responsibility is simple. Say one wanted to know who is responsible for a particular political decision, all one needs to do is find which representatives gave their assent to the relevant act—something that is usually a matter of public record. If one disagrees with their actions, they can express this by casting a vote against their representative in the next election.

This process though, is disrupted when technologies are used to make decisions as the power that is lent to the representative by their electorate is then ceded to the decision-making technologies: the actual seat of power thus shifts away from elected representatives towards the technologies that they use to make their decisions. The problem of the responsibility gap again arises—it does not seem that we can hold an elected representative responsible for the decisions made by technology, in much the same way that we cannot hold the operator or programmer responsible in the example above. Any attempt to ascribe responsibility can be sidestepped by an appeal to the autonomy of the technology, or an appeal to ignorance regarding the workings of the technology.

This said, there may be a way to ascribe responsibility in this context. It may be possible to force responsibility by way of "introducing new liability mechanisms" (Santoni de Sio and Mecacci, 2021., p. 17). Such mechanisms will largely consist of legal tools to *assign* responsibility for the decisions of such technologies to individuals and/or organisations. This perhaps could be done by way of ascribing those that *make use* of such technologies responsibility for their impact in a somewhat brute fashion, e.g., by entering a legal

arrangement that states that by choosing to deploy the technologies, you agree to be held responsible for their impact.

There are several issues with assigning responsibility in this way that undermines its effectiveness in a political context. Firstly, there is something of a conflict of interest present in doing this in a governmental context. As noted by Santoni de Sio and Mecacci, "liability regimes are managed by the State and require strict standards of causation, evidence and seriousness" (ibid., p. 18), and as such, assigning responsibility by way of legislation requires that the legislature passes an act that opens themselves or their peers up to being legally responsible for technologies they delegate decision making to. It is difficult to see such self-regulation passing, if not due to the inherent conflict of interest, then because of the second issue—put simply, there is a distinction between legal responsibility and actual culpability. If something goes awry due to the technologies a department utilises, then this is owed to the technology, not the elected representative that heads the relevant department. They may have no control over the technology, or otherwise be ignorant of its workings.

Thirdly, the opacity of these technologies makes putting together a case that meets the stringent legal standard of such legislation prohibitively difficult and laborious. If users are ignorant of the workings of such technologies, and the technologies themselves are convoluted and inscrutable, then it becomes hard to define the precise causal role played by the technology. As such it becomes prohibitively difficult to distinguish between instances where there is a poor case for legal liability—e.g., in cases of one-off malfunctions in the technology—and those where there is a robust case for legal liability, e.g., if the technology is improperly calibrated or operating with questionable data. As

such, assigning responsibility to legislators via legal decree is at best against the direct interests of the legislators who would be required to create the legislation, and at worst unworkable on a practical level.

8.6 THE RESPONSIBILITY GAP: A FEATURE, NOT A BUG

The difficulties in attributing responsibility to actors when humanlike technologies are introduced into governmental decision-making processes gestures towards a fatalistic understanding of technology that threatens our collective agency. The robustness of the responsibility gap allows thinkers to frame technologies as neutral, decontextualised and apolitical: the gap exists because these technologies belong to a space that sits outside of political responsibility and accountability, as they lack the basic features required of moral agents, i.e., "they do not have mental states and intendings to act" (Johnson, 2006., p. 202).[44] Moreover, thinking that we can find an individual to whom we can ascribe responsibility, to such an understanding of technology, mistakes the kind of relationship we have with technology.

With such an understanding, technology is not something we have dominion over (hence the responsibility gap), but rather it is something with its own autonomous logic, something that "follows a unilinear course, a fixed track, from less to more advanced configurations" (Feenberg, 2010., p. 8) and because of the "*epistemic* obligation" (Bjerring and Busch, 2020., p. 6) created by technology's supra-human capacities, "social institutions must adapt to the 'imperatives' of the technological base" (op cit., 2010., p. 9). In essence, technology, when understood in such a way, loses its instrumental character—it

[44] It is perhaps no surprise that this is the case, considering that materialism—the background impetus behind these technologies—is, as we saw in chapters 2 – 4, ill-equipped to deal with issues regarding human subjectivity.

ceases to be something that we do, or something that we use—and instead becomes "a way of revealing" (Heidegger, 1954., p. 318), something that operates primarily by influencing what shows up for us as useful (Brayford, 2020a., p. 530). As such, technology (to use Heideggerian terminology) "*challenges-forth*" (op cit., 1954., p. 335) a particular understanding of the world, and thereby effectively interprets the world in our place (Brayford, 2020b., p. 2). This technologically interpreted world brings with it "new potentialities for action" (ibid): once deployed to a society, technologies articulate a set of imperatives (op cit., 2020a., p. 530) that its social organisations have an obligation to follow.

In the context of using humanlike technologies to make governmental decisions or to optimise public services, these imperatives are literal. For instance, if we return to the example above regarding crime, we see that the algorithm's ability to use location data and crime reports to create a map of areas where the deployment of resources may be most efficacious can be fruitfully explained via this lens. Combining and analysing the data *reveals* to governments where best to allocate their resources—it discloses to them an already interpreted world, where certain geographical locations show up as *useful* for the project of combatting crime. As such, new potentialities of action are discovered, and the technologies issue an imperative to us, telling us what we need to do: e.g., the technologies tell us the most efficient patrol route so police forces are present in the times and places where they are most likely to be needed. In such instances, it is human actors that are held responsible by supposedly neutral and apolitical technologies. If the police forces were to disregard the imperatives handed to them by the technology, it could plausibly be construed as a dereliction of their duties. The same would also hold true for government ministers and other elected representatives—if they were to disregard the interpretation of

the world and the potentialities for action as revealed to them by the technologies they consult, then this too could be seen as a dereliction of their duties as public servants. This is because, as mentioned above, there is an epistemic obligation to follow the imperatives handed to us by technologies in cases where they possess supra-human abilities—to not follow their imperatives is not only a violation of this obligation but is also a violation of the democratic obligation for elected representatives to efficiently manage matters of public interest.

8.7 OBLIGATIONS AND AGENCY: THERE IS NO ALTERNATIVE

This understanding of technology, where imperatives passed to societal actors by apolitical technologies leaves governments in something of a difficult situation. By not following their epistemic obligation to follow the diktats of technology, they risk losing the benefits that such technologies can bring which may consequently bring about negative ramifications, not just in the realm of public service provision, but also electorally. But, on the other hand, if they *do* choose to respect their epistemic obligation and effectively cede decision-making to the perceived wisdom of technologies, then the collective agency of the electorate is diminished as a result: democracy—and by extension, our collective political agency—becomes moot. If technology is seen to exist outside of the realm of political responsibility and there is the pressure on elected officials to follow its diktats, then whom one chooses to vote for is of only nominal relevance—it does not matter what colour rosette our elected representatives wear if it is ultimately technology that makes the decisions. Relying on technologies in such a way puts the effective seat of power *beyond* democratic accountability: the electorate cannot choose to vote out the decision-making technologies and so democratic accountability is lost.

The effect that this has on collective agency is more profound than it may first appear. Putting power beyond the reach of the electorate does not just render elections moot, but it also encourages an obscuring of different political potentialities. Seeing technology as an autonomous and apolitical force that issues imperatives to a society is a distortion of technology's true relation with society, caused by a neglect of the concrete conditions from which technologies are developed and to which they are deployed. In the case of the distorted understanding of technology, "[technologies] are treated as 'forces-in-themselves', removed from the context of power relations in which they are constituted and which determine their use and function" (Marcuse, 1977., p. 168): they are mistakenly "reified [and] hypostatized as fate" (ibid). When reified as fate, technologies slip out from our dominion. Instead of helping us to achieve our aims, they set the political agenda via issuing imperatives to follow—their subordinate role is subverted and replaced by a role characterised by deterministic authority. By taking up this fatalistic stance towards technology, we risk losing sight of the fact that, because they are grounded in a particular context that structures their specific form and effects (op cit., 2020b., p. 4), technologies are *not* apolitical, but instead bear the hallmarks of the political context in which they are developed and deployed.

Technologies are an embodiment of a particular set of political values and imperatives[45]—they are "always a historical-social *project* … [insofar as they relate to] what a society and its ruling interests intend to do with men and things" (Marcuse, 1965., p. 168). The notion that they are autonomous and apolitical is thus incorrect; an aberration caused by a neglect of the concrete situations from which technologies emerge from and are deployed to. In

[45] The drive towards efficiency that precipitates technology's presence in business and governmental contexts is one such value here, as is the Promethean belief that all can be put to use in increasing quality of life.

actuality, technologies are a site of political manipulation (op cit., 2020b., p. 4)—a way of "political power assert[ing] itself" (Marcuse, 1964., p. 5). Technology therefore, when properly understood, is seen as a locus of political power that organises society in line with the political values embedded within them: they are a way of realising a *possible* societal configuration, informed by their embedded values and sympathetic to the wishes of those that create, deploy and make use of them.

Neglecting the political nature of our technologies causes a misunderstanding of the relationship between technology and society and it is this misunderstanding that constitutes the greatest danger to our political agency. As our technologies always belong to a socio-political context (insofar as they embody its values) and are a site of political power (insofar as they inform a particular, *future*, socio-political potentiality) technologies and the society to which they are deployed sit in a dialectical relationship with one another. Neglecting this and incorrectly viewing technology as in possession of apolitical autonomy means that the dialectic relationship between society and technology, where each influence the other, is misrepresented as a relationship characterised by unilateral influence, of technology on society (op cit., 2020a., p. 531). By losing sight of the fact that technologies are inherently political and treating them as if they occupy a space beyond human control, we relinquish whatever influence it is that we may have had over them, thus giving them free reign over shaping future socio-political potentialities. Society and its politics thus risks becoming atrophied. We become unable to reckon with different societal potentialities because our political agency—i.e., the ability to shape our own political situations in a manner sympathetic to our wishes—becomes lost amongst the imperatives handed to us—or perhaps more accurately, handed to our elected representatives—by decision-making technologies. By denying ourselves the ability to realise alternative

political potentialities, we condemn ourselves to whatever future potentiality becomes realised: all alternatives become obscured.

8.8 THE IRONY REVISITED

From this we can see how the deployment of human technologies—those made to emulate our image, to meet or exceed human capacities—can contribute to the great irony mentioned at the start of this chapter. The Promethean wish of bringing everything under our control through the power of technology so we can put it to task in solving our problems reaches its zenith in the widespread undermining of agency. If humanlike technologies saturate the workplace, we risk entering into a situation where meaningful work is put beyond of the reach of much of the populace, thereby interfering with their ability to form a robust and rewarding sense of identity which, in turn, undermines their potential to live a life that is expressive of their own agency. Similarly, if such technologies become incorporated into governmental decision-making processes, because of the epistemic obligation for elected officials to follow the imperatives of the technologies they consult, we risk entering into a situation where political expression becomes sanitised, mollified and disconnected to the actual loci of political power. Elections, and other engagements with political life, become futile as our socio-political contexts become informed by the diktats of technology: political change thereby is rendered impossible.

Additionally, looking at the deployment of our technologies through the lens of agency explains, with considerable persuasiveness why the deployment of humanlike technologies is valuable in some areas, and unattractive in others. For instance, our conclusion that humanlike technologies are of positive social utility in the arenas of healthcare (§6.3) and finance (§6.6 and §6.7) and of negative social utility in warfare (§7.4 – §7.8), the

workplace (§8.4 and §8.5) and politics (§8.6 and §8.7) can be explained by an appeal to agency. Put simply, the ceding of agency is permissible in healthcare and finance because, if successful, it is likely to beget gains in agency greater than that lost (healthy and wealthy individuals are better able to engage with their agency than the poor and the sick) whereas the ceding of agency in the other fields risks a net reduction of agency: a civilian or a soldier killed by an AWS, an individual who cannot find work, and an elector whose vote is worthless, share in common a reduction in agency. Indeed, this may have been the impetus behind the genesis of the Promethean attitude, that in turn helps to explain *why* materialism became part of the prevailing doxa. In the context of the early period of industrialisation, where life was characterised by poor living conditions, subsistence incomes and widespread ill health, the Promethean promise of a healthier and wealthier life would have held irresistible appeal. The flaws of materialism, and the negative impacts of its technologies were lost behind this irresistible appeal and were carried through to the modern day primarily via inertia. That materialism and its technologies threaten to jeopardise our agency has little truck in those overwhelmingly concerned with meeting their immediate concrete needs: one does not worry about their agency when they are concerned about where their next meal is coming from. Yet for many, especially those in developed nations, the concrete conditions of their lives have changed: our lives are largely free from the poverty that was common in the past. Technology has enabled us to live a life of considerably greater ease than those lived by our ancestors. But with this improvement in living conditions, our attitudes should shift—the threat of losing our agency somewhere amongst the incessant drive to raise our living standards further should be that which concerns us now that we have largely escaped the extreme poverty of the past.

As noted above, it is perhaps unsurprising that the Promethean attitude contains within it a disregard for matters of agency because of its materialist sympathies. Recall, as we saw in chapters 2 – 4, that materialism's faith that the *view from nowhere* is the methodology *par excellence* in uncovering the ultimate truth about the world engenders difficulties in accounting for matters of human subjectivity: from the Archimedean point, the image of what it is to be human is impoverished and the important features of human life—our consciousness, our free will and our selfhood—are neglected. By endorsing materialism, even tacitly, those technology developers and users who adhere to the Promethean attitude, become unable to properly understand what it is that makes us human, and consequently they risk inadvertently creating and deploying technologies that violate our humanity. The errors of materialism become translated into our technologies and therefore, our agency is ceded to technologies and our selfhood is reduced to a mathematical variable—a particular weighting in a complex and incomprehensible equation. What is more, those who adhere to the Promethean attitude are unable to grapple with the potential effects of this, as their enthusiasm for their technologies—their faith in the power of technology to bring nought but good to humanity—trammels their ability to see the potential negative effects of the technologies that they are surrounded by. A lack of healthy scepticism towards technology, if we are not careful, will lead to the creation of a deterministic, formulaic and bureaucratic world, where our material needs are perhaps catered for, but at the expense of violating that which sets us humans apart from the natural world—our ability to create ourselves.

Insofar as this is true, and the features of our subjectivity risk being side-lined by the Promethean wish to put our world to task, we must question the social utility of the technologies that are supposedly created in our image. If these technologies are not treated with appropriate scepticism, then the false image of humankind that is promulgated by

materialism—i.e., that of an unconscious and selfless configuration of matter at the mercy of physical laws—will be the image of humanity that our technologies will come to serve. The mistakes of materialism will become manifest in our lived existence as the technologies strive to create a world fit for the aberrated, impoverished image of humankind. We will be richer than ever before, and our lives will be longer and our ailments fewer—but this comes at the price of losing meaningful freedom. Our working lives will be reduced to whatever is left over after automation. Elections will become pointless, a superficial choice between whichever politician seems to best align with our ideological leanings—after all, all the real decisions will be made by technology on their behalf. That which is unique to humans will come into conflict with a world that neglects the fundamental features of our existence: our consciousness, our freedom and our selfhood will have to contend with technological forces which pay them no heed. Our ability to live a life of our own will become—perhaps irreparably—damaged. The exercise of freedom in this future will be reduced to impotent fancy: one can imagine that the most meaningful choices one would make in such a future pertains to the brand of refrigerator they wish to purchase, or what style of shoe they most prefer. Once the novelty of this begins to wane, we will find ourselves fundamentally dissatisfied, racked by the nagging anguish and misery of living an artificial and inauthentic life—a life that we have only the most superficial control over. The danger that this future becomes realised is made all the more likely should we not recognise that it is humankind—not our technologies—that is in charge. If we allow the optimism of the Promethean attitude to obscure the dangers of technology, we risk surreptitiously contributing to this situation. Instead of using technology to our advantage, our misunderstandings of what we are and what power our technologies possess threaten to contribute to our subjugation at the hands of an *unhuman* technological bureaucracy. The Promethean attitude may make some look forward to the

future with relish, but the effects of the materialism and technological optimism incorporated into the attitude suggest that once we get there, we may not like what we are forced to reckon with.

9 CONCLUDING REMARKS

In part one of this thesis, I began by noting that a widespread belief in materialism, in the context of ever-increasing technological power and large flows of investment, has reignited the faith that we will be able to create technologies in our image. The ultimate goal of this faith is to create strong-AI—technologies that are replete with human subjective faculties. If technologies are to be made in our image, it is important to understand *what* image of humanity guides their development. To isolate this image, I gave a brief historical sketch of the image. It emerged from the failings of the dualism espoused by Socrates, Avicenna and Descartes and the image of humankind—i.e., that of humans as a composite of mind and body—that dualism advances. Such an image is flawed—as Gassendi was keen to highlight—because it is unable to offer an account of how the mental can have any impact on the material: it is difficult to see how two radically different substances can interact with one another. Materialism's solution to this problem is to endorse monism: all worldly phenomena, from the falling of autumn leaves to eudemonic flourishing, to the materialist, is owed in some way to matter. I then noted that this necessitates a materialist analysis of the mind. I then isolated illusionist machine functionalism (IMF)—which holds both that the mental is best defined functionally, and that consciousness is ultimately illusory—as not only the most natural and plausible theory of mind for a consistent materialist to endorse, but also the one most closely entwined with the project of making technology in our image. I then noted that IMF is bolstered by materialism's hegemonic position. I argued that materialism's place in the prevailing doxa cannot be owed to IMF, because materialism lends support to IMF, not vice versa, and nor can it be owed to its dominance in academic philosophy because the academy has not been terribly influential on ordinary people.

With this in mind, I then argued that the popular appeal of materialism is owed primarily to the fact that it hooks into another external phenomena that then acts to imbue it with legitimacy in a by-proxy fashion. Looking at the societies with the greatest assent to materialism—the capitalist west and the once communist east—allowed me to ascertain what this external phenomenon is. I advanced the idea that, in capitalist societies, the dominance of materialism is owed primarily to the fact that it removes the theoretical limit of putting the natural world to task in improving our quality of life. Technology allows for the manipulation of natural resources, which in turn drives economic growth, which then improves the quality of life of citizens. Because technology operates via manipulating the material world, materialism's insistence that all phenomena are owed to matter means that technology is able, in theory, to manipulate *any* phenomena to generate gains in quality of life. The communist story, as we saw, is slightly different. Karl Marx's response to the Hegelian understanding of historical progress helped to advance a materialist conception of history. This was then taken up by Marx's patron, Friedrich Engels, who sought to demonstrate that the entirety of the natural world operated in concordance with dialectical laws. I argued that, by doing this, Engels expanded the scope of Marx's materialism and made it totalising—thus introducing an explicit element of determinism in the unfolding of history. This deterministic element was seized upon by Vladimir Lenin and Joseph Stalin and used to justify their revolutionary activities—their overthrow of Tsarist Russia and the founding of the Soviet Union was the fulfilment of a prophecy writ large upon the supposed laws of the natural world. I suggested that this helped to secure assent to materialism in two ways. Devout communists saw materialism (and thus the revolution) as a way of securing a better quality of life without first enduring the hardships of capitalist development. Dissenters, on the other hand, had their assent secured through oppressive

means—deviation from a materialist world view was viewed as an example of anti-revolutionary sentiment and was suppressed accordingly.

The commonality between the capitalist justification of materialism, and the communist justification of materialism—i.e., the implicit belief that materialism plays a role in bringing about a higher standard of living—is that which legitimises materialism. What is more, this belief in materialism's value in bringing about a better future encourages a sense of technological optimism. These two factors, I argued, find their synthesis in the *Promethean attitude*—i.e., the belief that anything that shows up to us as a problem is surmountable, provided that we have the correct technologies to help us. I suggested that this attitude has been a useful one to hold—it *has* helped us to achieve a standard of living that outstrips that found only a few generations ago—but this does not mean that it will *always* be a useful attitude to hold. At this point, I called for an evaluation of the Promethean attitude: this was split into two parts. The remainder of part one focussed on an evaluation of the materialism associated with the attitude—we returned directly to an evaluation of the Promethean attitude in part two.

In chapter 2, I advanced the thesis that the primacy of the deficient mode of Care and a staunch subservience to the *view from nowhere* leads to an impoverished image of what it means to be human: our subjectivity becomes lost behind the ways in which we are object-like. Our consciousness, unable to be accounted for with the limited conceptual toolkit of particles, forces and waves, is thus held by IMF to be illusory—if materialism is true then the most plausible option we have to account for our consciousness is to deny it even exists. The illusionist thus holds that we trick ourselves into thinking that we are conscious: our mental lives are nought but a clever linguistic hoax—a product of a

narrative that hooks into the material goings on that supposedly underlie our illusory conscious experience. But, as I argued, such illusionist conclusions are not a reflection of what is actually the case. Rather, they are the product of a misapplied methodology and an improper faith in the notion of *givenness*. The suppression of our subjective faculties that we perform when seeking the view from nowhere causes aberrations in the images that they reveal. Baked into this *view from nowhere* is a flaw—materialism forces us to account for matters of human subjectivity through a methodology that *explicitly suppresses* human subjectivity. It is therefore unsurprising that IMF reaches its illusionist conclusions—its methodology encourages its adherents to be hostile towards or suspicious of the very subjectivity that they seek to analyse.

In chapter 3, I argued that the same also broadly holds true for free will. As we saw, by operating from a position of detached scientific neutrality both arguments seeking to account for free will and seeking to demonstrate determinism fail. We saw this in Helen Steward's attempt to account for free will via her Agency Incompatibilism. The *view from nowhere* mandated by materialism collapses the distinction between humans and the rest of the animal kingdom. Seen from the outside, humans and animals are surprisingly alike—we both, in Steward's view, are capable of arresting the course of determined events to give us space to exercise our free will by the sheer fact that we are self-moving. But, as I suggested, this causes another aberration to the image of humans—our free will becomes marred by reaction and pettiness. We are free, under such a view, only insofar as we can choose to respond to that which has befallen us, and even then, this freedom is one that is only found in a choice of *technique*: we cannot choose what we have to do, but we can choose how we do it. Benjamin Libet's experiments that purport to show that our freedom is at best, a veto, or at worst, wholly illusory, present a similarly aberrated image. This was

owed to the fact that both theories choose to hold an *atomistic unit* of an action to be a suitable analogue of an action proper. Forced into doing this by their materialist commitments, they lose sight of the existential significance of our actions and thus misunderstand what it means to be free, thereby impoverishing the image of humanity further.

Chapter 4 argued that selfhood is also missed out from materialism's impoverished image of humanity. I explored three options available to the materialist in accounting for selfhood before showing them to be unattractive or incoherent. I began with Peter Van Inwagen's attempt to account for our selfhood by finding a part of the material world that is numerically identical to it (i.e., the brain), before arguing that such a position is marred by its unattractive conclusions: we come to believe that our self can be sustained by an unembodied, necessarily retrospective brain wired into a life support system. The second option explored—advanced by Daniel Dennett—holds that, like our consciousness, our selfhood is another linguistic illusion: we are the *I* in our own story about ourselves. But here little improvement is found. Presenting our selfhood as a centre of narrative gravity means that we do not do justice to our phenomenal consciousness—fictional entities cannot *feel*—and moreover, it suggests an incoherency: thinking of a self as a centre of narrative gravity is an act of self-interpretation, and such, suggests the existence of non-fictional selves. A second illusionist option was examined—that found in the work of Thomas Metzinger. His view is that selves are an illusion caused by a representation of our bodies, in a representation of our environments, realised on a neuro-biological level. We are unaware of our representational status because our brains are quick at amending the representations and evolving the ability to recognise them would have been too costly in terms of energy. Moreover, the possession of these representations, in Metzinger's view,

allow us the ability to *do* things—our actions are contingent on such representations. Yet, I argued that Metzinger's arguments against naïve realism hold no water, and his assertion that action relies on representation is refuted by non-representational theories of action—in short, there is no need to duplicate in a representation what is already out in the world for us to use. As such, there are no reasons to endorse Metzinger's illusionist conclusions.

I argued that this impoverished image of humanity scuppers the project of realising a strong-AI. With the impoverished image as the guide to humanlike technologies, the best technology developers can create under the auspices of the materialist paradigm is weak-AI—an AI that is humanlike insofar as it *emulates* human capacities. Yet, I assumed that there are those who would not allow the impoverished image of humanity to interfere with their faith in the creation of a strong-AI. To undermine these, I explored what their justifications for assuming an AI to be in possession of human subjective faculties could be. I identified two possible justifications. Behaviourism, which is undermined by the fact that behaviour is no infallible indicator of subjective human capacities (as evidenced by John Searle's Chinese room thought experiment) and materialism. I advanced the thesis that materialism is too vague in definition to be a sufficient justification. To do this, I explored Daniel Stoljar's comprehensive survey of potential definitions before underscoring his insights with an appeal to Hempel's Dilemma. Put briefly, calling oneself a materialist is to endorse either an incorrect metaphysic—as our physical theories will certainly progress—or an unknowable metaphysic—if, in endorsing materialism, one has a finished, ideal physical theory in mind, then they have no way of knowing what this entails, owing to the fact that it belongs to the future. As such, I contended that we have no good reasons to endorse materialism—not only does it present an impoverished image of humanity, but it does so under the auspices of a metaphysic that is either wrong, or

completely opaque in content. As such, I argue that the belief that strong-AI is possible in a materialist paradigm is either motivated by a naïve faith in the prevailing doxa, or otherwise, is the collateral product of an endorsement of materialism made for a reason that sits beyond materialism in and of itself. I suggested that this reason is the belief that a materialist metaphysic is a useful adjunct to the Promethean attitude and its promise of a better standard of living.

Here I began part two, where I argued that the Promethean promise of a better life is ultimately empty: technology is poised to make us healthier and wealthier, but it is also likely to make our wars less ethical, bloodier and more destructive, and diminish our individual and collective agency, thereby leading us to a future of profound misery.

In chapter 6, I explored the likely impact of these humanlike technologies on healthcare provision. It was argued that the Promethean attitude and its technological products are likely of positive social utility in this field. Should our technologies become able to parse the totality of medical literature, then healthcare provision is poised to become more accurate and effective, whilst simultaneously becoming easier to access, and less burdensome for medical practitioners. This benefit is offset somewhat by the foibles of such technologies, specifically, their lack of appreciation for unusual or rare clinical presentations of disease and their potential for misdiagnosis should they be trained with low quality data, or their modelling of the patient be somehow erroneous, but I suggested that this can be mitigated by human oversight of the technologies. The opaque *black-box* nature of the technologies hinders, I argued, direct oversight—their complexity stands in the way of their comprehensibility—but I then suggested that this is able to be mitigated by rethinking the relationship between the technology and the medical practitioner. If the

practitioner violates their epistemic obligation to defer diagnoses to the technologies by remaining sceptical of their suggestions and instead thinks of them as a *tool* to complement their own diagnostic endeavours, then their incomprehensibility is somewhat irrelevant—a medical practitioner does not need to understand the intricacies of the equipment they used for it to be of use.

Next, I argued that the benefits of humanlike technologies in healthcare are mirrored by the benefits that they can bring in finance. The use of humanlike technologies, I claimed, allow lenders to better understand the risks of lending to those with weak or non-existent credit profiles through analysing data not traditionally consulted in credit-profiling. This allows access to financial products to be extended. What is more, because of their impartiality they can also provide more equitable access to financial products for those who may, for whatever reason, be discriminated against by a lender. I also explored how automated trading algorithms are able to better leverage arbitrage opportunities in the market, allowing those with investments to grow their assets more quickly and with higher returns. To temper this optimism, I discussed two problems unique to automated trading algorithms: the danger of duelling automated traders, and the danger that such technologies impede financial understanding, and thus impact negatively on oversight. The first danger, I argued, can be mitigated by both developing the technologies to be sufficiently adaptive to changes in market regime, and developing them with simplicity and comprehensibility in mind. Doing this allows greater epistemic access to the technologies thereby aiding the isolation and eradication of any unsatisfactory behaviours. I also suggested that designing these technologies with simplicity and comprehensibility in mind can mitigate the second danger. The concern is that these technologies threaten to make the already fairly opaque—finance and economics—more so, thereby affording unscrupulous actors freedom from

oversight: something further complicated by the difficulty in attributing responsibility for the actions of the technologies. I argued that increased comprehensibility via simplicity also helps to assuage this issue, and as such, the attitude and its humanlike technologies are of positive social utility here also. Although, as we saw in chapters 7 and 8, these gains in health and wealth are offset by the fact that these technologies bring about misery and anguish elsewhere.

In chapter 7, I argued that the gains technology may have with regard to health and wealth are offset by their potential to intensify misery in the context of war: the deployment of humanlike technologies—specifically, automated weapons systems (AWS)—to the battlefield will be of negative social utility as these technologies are poised to make warfare less ethical and more dangerous and destructive. I began by noting that adherents to the Promethean attitude are seeking to develop truly autonomous weaponry—i.e., that which does not have human oversight or direction—for deployment on the battlefield. I drew upon Thomas Nagel's idea that warfare should be conducted with respect towards one's opponents' subjectivity in mind to evidence widespread distaste towards the idea of AWS: losing one's life at the behest of an AWS violates this principle. I advanced the discussion by exploring Ronald C. Arkin's enthusiasm for AWS—he suggests that such technologies are able to wage war more ethically than human soldier could, owed to their supra-human capacities, their expendability and their lack of emotion. I temper Arkin's enthusiasm by offering two caveats to his position from the literature—AWS must be developed to respect the laws of war and must have adequately accurate targeting capacities before they can be deployed. I then advance several arguments against AWS. I began by arguing that the supposed benefits of AWS remove the necessity of deploying human personnel to the battlefield and this, when combined with the expendability of

AWS undermines the ability of a force to achieve their aims whilst leaving the impetus for prosecuting war intact. In short, one cannot impose one's will on another by destroying that which they do not care about. As such, I argued that the deployment of AWS could lead to an absurd opening spectacle before a traditional human-vs-human, or a human-vs-AWS war commences. I then took the latter to argue that the supposed ethical benefits of deploying AWS are undermined by the fact that, in a battlefield with both AWS and human soldiers, the increased stresses borne by soldiers facing a profoundly lethal adversary will likely cause *more* unethical behaviours, as their stresses translate into ethical transgressions. I then investigated the intricacies of target selection on a battlefield—i.e., the distinction between legitimate and illegitimate targets stipulated by the laws of war—before noting its Heideggerian nature. My argument here was that targeting requires the AWS to grapple with the ontic-ontological significance of battlefield entities, but it cannot do so because it lacks the conditions necessary for Care—i.e., consciousness, free will and selfhood. As such, the deployment of AWS would likely have a *negative* impact on warfare, and thus the Promethean attitude and its humanlike technologies are of poor social utility here also. By making warfare less ethical and bloodier, these technologies threaten to interfere with the ability of those embroiled in conflict to live a free and authentic life of their own choosing: a civilian or a soldier living in fear of an AWS, or worse, that has been killed by an AWS has their agency violated at the behest of technology.

In the final chapter of the thesis, I explored the irony of the Promethean attitude: that by trying to control our world via humanlike technologies, we risk diminishing our own agency, thus condemning more people to a life of nagging misery and anguish. This, in my view, is owed not only to the fact that materialism (as I explained in chapters 2 – 4) is ill-

equipped to deal with matters of human subjectivity but also to the fact that the Promethean attitude's optimism toward technology conceals this fact. Adherents to the Promethean attitude cannot see the potential negative effects of its associated technologies if they believe technology to be an overwhelming good. To evidence this irony, I looked at how the deployment of humanlike technology to the workplace interferes with the human capacity for self-creation, by way of reducing the likelihood that a person finds themselves in meaningful employment. The argument here was twofold. Firstly, as technologies displace human workers, meaningful work (as delineated by Andrea Veltman) becomes scarcer as the gross availability of *all* jobs—edifying or otherwise—declines. Secondly, the work that *would* remain available is less likely to be edifying, as it is likely to be reduced in scope to meaningless oversight of capable technologies in order to sate some superfluous moral, legal or political principle. These two arguments culminated in the suggestion that the deployment of humanlike technologies results in a decrease in human agency. This conclusion was also found in the case that technologies become incorporated in governmental decision-making processes. I argued that this leads, effectively, to a situation where the actual seat of political power shifts from our elected representatives to the technologies they utilise to guide their decisions. In my view, this shift in the locus of political power arises from a misunderstanding of the proper nature of our technologies. By understanding our technologies as neutral, apolitical and deterministic, we lose sight of their inherently political nature: embedded within technologies are political principles, that then go on to inform the political culture of the societies to which they are deployed. By heeding the epistemic obligations of the technologies—whether for political reasons, operational reasons, or otherwise—our elected officials allow this improper view of technology to interfere with our collective political agency. Two negative consequences arise from this shift. Firstly, elections are rendered moot—if the technology is the true

bearer of political power, whom one votes for is irrelevant. Secondly, political culture, through the closing off or obscuring of different political potentialities, becomes solidified and atrophied. I ended by affirming my position that ultimately, these materialist technologies are poised to make us healthier and wealthier, whilst simultaneously stripping us of the ability to *do* much of worth with our lives, thereby trapping us into a position characterised by impotence and woe.

From this we can see that the utility of the Promethean attitude is beginning to wane. The danger it presents to us is profound—it blinds us not only to the fact that materialism is justified not by rationality, but by its place in the prevailing doxa, but also to the potential misery that the technologies that are developed under its auspices threatens to bring. Lured by utopian promises, we risk creating an inhospitable situation for ourselves. Literally, in the case of the climate crisis it has already precipitated, and figuratively, in the sense that it has prepared for us a future which will not respect the subtleties of humanity—the very subtleties that materialism encourages us to think of as illusory: our consciousness, our free will and our selfhood. I do not wish to suggest that we should tear the Promethean attitude, root and branch, from our thinking. Such would be not only impossible but of questionable attractiveness as it would also strip from us a sense of hope that is vital to avoiding the trap that we have set for ourselves. But I do think that we should excise the failings of the attitude and allow our Promethean visions to be guided by an adequate understanding of what it means to be human. By this, I suggest that materialism and the impoverished image of humanity that it creates should be allowed to fade into the annals of philosophical history. The Promethean attitude is correct when it suggests that we can make a future that is better than the past, but should it retain its materialist associations, its products will be marred as a result. The same also holds true for its unwavering adherence to technological

optimism—if we are unable to anticipate the potential negative effects of our technologies, then instead of correcting course to avoid them, we will be forced to reckon with them once they are already here: a task made all the more difficult by the magnitude of technological power, and the opacity of its workings. Yet if this is done, and we allow the Promethean attitude to quietly shed its materialist baggage, we can return to hoping for a better future. Free from materialism's anti-humanist conclusions, we would be better equipped to understand the true utility and significance of our technologies. We would be able to create a world, not where selves are denied, but where selves can flourish—where our freedom is not suppressed by our technologies but is bolstered by them: a world created in our image.

10 REFERENCES

Aristotle, n.d./1995. *Politics*. Translated from Ancient Greek by E. Baker. Oxford, UK: Oxford University Press.

Aristotle, n.d./2004. *Nicomachean Ethics*. Translated from Ancient Greek by R. Crisp. Cambridge, UK: Cambridge University Press.

Arkin, R. C., 2010. The Case for Ethical Autonomy in Unmanned Systems. *Journal of Military Ethics*. 9(4). pp. 332 – 341.

Athalye, A., Engstrom, L, Ilyas, A., and Kwok, K., 2018. Synthesizing Robust Adversarial Examples. *Proceedings of the 35th International Conference on Machine Learning*. 80. pp. 284 – 293.

Aziz, S. and Dowling, M., 2019. Machine Learning and AI for Risk Management. In: *Disrupting Finance: FinTech and Strategy in the 21st Century*. T. Lynn, J. G. Mooney, P. Rosati and M Cummings (eds). London, UK: Palgrave MacMillan. Ch. 3.

Barber. J., 1979. The Establishment of Intellectual Orthodoxy in the U.S.S.R. 1928-1934. *Past and Present*. 83(1). pp. 141 – 164.

Bartlett, R., Morese, A., Stanton, R., and Wallace, N., 2019. Consumer-Lending Discrimination in the FinTech Era. *UC Berkeley Public Law Paper*. [pdf] Available at: <https://dx.doi.org/10.2139/ssrn.3063448>.

Beebee, H., 2014. Radical Indeterminism and Top-down causation, *Res Philosophica*. 91. pp. 537-545.

Bjerring, J. C. and Busch, J., 2020. Artificial Intelligence and Patient-Centred Decision-Making. *Philosophy and Technology*. [online] Available at: <https://doi.org/10.1007/s13347-019-00391-6>.

Block, N., 1978. Troubles with Functionalism, In: N. Block, ed. 1980. *Readings in Philosophy of Psychology: Volume One*. London, UK: Methuen. pp. 268 – 305.

Bocheński, J. M., 1967. Why Studies in Soviet Philosophy? In: *Philosophy in the Soviet Union: A Survey of the Mid-Sixties*. E. Laszlo (ed). Dordrecht, NL: D. Reidel Publishing Company. pp. 1 – 12.

Borch, C., 2016. High-frequency trading, algorithmic finance and the Flash Crash: reflections on eventalization. *Economy and Society*. 45(3-4). pp. 350 – 378.

Bostrom, N., 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford, UK: Oxford University Press.

Botvnick, M. and Cohen, J., 1998. Rubber hands ‘feel’ touch that eyes see. *Nature*. 391. p. 756.

Brayford, K. M., 2020a. Myth and Technology: Finding Philosophy's Role in Technological Change. *Human Affairs*. 30. pp. 526 – 534.

Brayford, K. M., 2020b. How to Live a Life of One's Own: Heidegger, Marcuse and Jonas on Technology and Alienation. *Philosophy and Technology*. [online] Available at: <https://doi.org/10.1007/s13347-020-00417-4>.

Captain, S., 2017. Can IBM's Watson do it all? *Fast Company*. [online] Available at: <https://www.fastcompany.com/3065339/can-ibms-watson-do-it-all>. Accessed 8 March 2021.

Carpenter, C., 2013. US Public Opinion on Autonomous Weapons. [pdf]. Available at: <https://www.duckofminerva.com/wp-content/uploads/2013/06/UMass-Survey_Public-Opinion-on-Autonomous-Weapons.pdf> Accessed 15 April 2021.

Challen, R., Denny, J., Pitt, M., et al., 2019. Artificial Intelligence, Bias and Clinical Safety. *British Medical Journal: Quality and Safety*. 28. pp. 231 – 237.

Chang, H., 2014. *Economics: The User's Guide*. London, UK: Pelican.

Chen, M., 2019. A Tale of Two Deficits: Causality and Care in Medical AI. *Philosophy and Technology*. 33. pp. 245 – 267.

Chomsky, N., 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Clark, A. and Chalmers, D., 1998. The Extended Mind. In: A. Clark., 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford, UK: Oxford University Press. pp. 220 – 232.

Clark, A., 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford, UK: Oxford University Press.

Clausewitz, C. von., 1832/2008. *On War*. Oxford, UK: Oxford University Press.

Coeckelbergh, M., 2015. The Invisible Robot of Global Finance: Making Visible Machines, People, and Places. *Computers and Society*. 45(3). pp. 287 – 289.

Cohen, G. A., 2000. *Karl Marx's Theory of History: A Defence*. Princeton, NJ: Princeton University Press.

Crane, T., 2003. *The Mechanical Mind: A philosophical introduction to minds, machines and mental representations*. 2nd Edition. London, UK: Routledge.

Daly, C., 1998., What are physical properties? *Pacific Philosophical Quarterly*. 79. pp. 196 – 217.

Davis, M., Kumiega, A., and Van Vliet, B., 2013. Ethics, Finance, and Automation: A Preliminary Survey of Problems in High Frequency Trading. *Science and Engineering Ethics*. 19. pp. 851 – 874.

de Boer, K., 2009. Hegel's Account of the Present: An Open-Ended History. In: *Hegel and History*. W. Dudley (ed). 2009. New York, NY: SUNY Press.

Demirgüç-Kunt, A., Klapper, K., Singer, D., et al., 2018. *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution.* Overview Booklet. Washington, DC: World Bank.

Dennett, D. C., 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge, MA: MIT Press.

Dennett, D. C., 1991a. *Explaining Consciousness*. London, UK: Penguin.

Dennett, D. C., 1991b. Real Patterns. *The Journal of Philosophy*. 88(1). pp. 27 – 51.

Dennett, D.C., 1992. The Self as a Center of Narrative Gravity. In: F. Kessel, P. Cole and D. Johnson (eds) *Self and Consciousness: Multiple Perspectives*. [online] Available at: <http://cogprints.org/266/1/selfctr.htm>. Accessed 10 October 2020.

Dennett, D. C., 2003. Who's on first? Heterophenomenology Explained. *Journal of Consciousness Studies*. 10(9). pp. 19 – 30.

Dennett, D. C., 2017. *From Bacteria to Bach and Back*. London, UK: Penguin.

Descartes, R., 1637. Discourse on the Method. In: *The Philosophical Writings of Descartes*: Volume One. 1985. Translated from French by J. Cottingham, R. Stoothoff and D. Murdoch. New York, NY: Cambridge University Press. pp. 111 – 152.

Descartes, R., 1641. Meditations of First Philosophy. In: *The Philosophical Writings of Descartes*: Volume Two. 1985. Translated from French by J. Cottingham, R. Stoothoff and D. Murdoch. New York, NY: Cambridge University Press. pp. 1 – 62.

Dinstein, Y., 2007. *The Conduct of Hostilities under the Law of International Armed Conflict*. Cambridge, UK: Cambridge University Press.

Domingos, P., 2012. A Few Useful Things To Know About Machine Learning. *Communications of the ACM*. 55(10). pp. 78 – 87.

Dowell, J. L., 2006. The Physical: Empirical, not Metaphysical. *Philosophical Studies*. 131. pp.25-60.

Dreyfus, H. L., 1991. *Being-in-the-World: A Commentary on Heidegger's 'Being and Time, Division I'*. Cambridge, MA: MIT Press.

Dreyfus, H. L., 2002. Intelligence without representation – Merleau-Ponty's critique of mental representation: The relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences*. 1. pp. 367 – 383.

Dryzek, J. S., 2013. *The Politics of the Earth: Environmental Discourses*. 3rd ed. Oxford, UK: Oxford University Press.

Eagleton, T., 2016. *Materialism*. New Haven, CT: Yale University Press.

El-Bizri, N., 2017. Avicenna. In: S. Leach and J. Tartaglia, eds. 2016. *Consciousness and the Great Philosophers.* Oxford, UK: Routledge. pp. 45 – 54.

Engels, F., 1878/1954. *Anti-Dühring: Herr Eugen Dühring's Revolution in Science.* Moscow, RU: Foreign Languages Publishing House.

Engels, F., 1883/1946. *Dialectics of Nature*. Translated from German by C. Dutt (ed). London, UK: Lawrence and Wishart.

Feenberg, A., 2010. *Between Reason and Experience: Essays in Technology and Modernity*. Cambridge, MA: MIT Press.

Ferraris, M., 2014. *Manifesto of New Realism*. Translated from Italian by: S. De Sanctis. Albany, NY: SUNY Press.

Flew, A., 1951. Locke and the Problem of Personal Identity. *Philosophy*. 26(96). pp. 53 – 68.

Ford, M., 2015. *The Rise of the Robots: Technology and the Threat of Mass Unemployment*. London, UK: Oneworld Publishing.

Frankish, K., 2016. 'Illusionism as a Theory of Consciousness' *Journal of Consciousness Studies*. 23(11-12). pp. 11 – 40.

Frege, G., 1918/1956. The Thought: A Logical Inquiry. *Mind*. 65(259). pp. 289 – 311.

Gallagher, S., 2017. *Enactivist Interventions: Rethinking the Mind*. Oxford, UK: Oxford University Press.

Gibbs, S., 2014. 'Google buys UK artificial intelligence startup Deepmind for £400m' *The Guardian*, [online] (Last modified on Thu 30 Nov 2017 18:19 GMT). Available at: <https://www.theguardian.com/technology/2014/jan/27/google-acquires-uk-artificial-intelligence-startup-deepmind>. Accessed: 4 December 2017.

Gibson, J. J., 1979/2015. *The Ecological Approach to Visual Perception: Classic Edition*. New York, NY: Psychology Press.

Goodman, L. E., 1992. *Avicenna*. London, UK: Routledge.

Goos, M., Manning, A. and Salomons, A., 2014. Explaining Job Polarization: Routine-Biased Technological Change and Offshoring. *The American Economic Review*. 104(8). pp. 2509 – 2525.

Guriev, S., 2018. Economic Drivers of Populism. *AEA Papers and Proceedings.* 108. pp. 200 – 203.

Haggard, P., and Magno, E., 1999. Localising Awareness of Action with Transcranial Magnetic Stimulation. *Experimental Brain Research*. 127. pp. 102 – 107.

Hansen, K. B., 2020. The Virtue of Simplicity: On Machine Learning Models in Algorithmic Trading. *Big Data and Society*. [online] Available at: <https://doi.org/10.1177%2F2053951720926558>.

Harman, G., 2005. *Guerrilla Metaphysics: Phenomenology and the Carpentry of Things.* Chicago, IL: Open Court.

Harman, G., 2011. *The Quadruple Object*. Winchester, UK: Zero Books.

Harnad, S., 1982. Consciousness: An Afterthought. *Cognition and Brain Theory.* 5. pp. 29 – 27. [online] Available at: <http://cogprints.org/1570/1/harnad82.consciousness.html>. Accessed 16 July 2020.

Hegel, G. W. F., 1807/1977. *Phenomenology of Spirit*. Translated from German by A. V. Miller. Oxford, UK: Oxford University Press.

Hegel, G. W. F., 1840/1988. *Introduction to The Philosophy of History*. Translated from German by: L. Rauch. Indianapolis, IN: Hackett.

Heidegger, M., 1927/2010. *Being and Time*. Translated from German by: J. Stambaugh. Albany, NY: SUNY Press.

Heidegger, M., 1954. The Question Concerning Technology. Translated from German by W. Lovitt. 1977. In: *Basic Writings: Revised and expanded edition*. 1993. D. F. Krell (ed). Oxford, UK: Routledge.

Heidegger, M., 1969/1972. *On Time and Being*. Translated from German by: J. Stambaugh. New York, NY: Harper & Row.

Hellman, G., 1985. Determination and Logical truth. *The Journal of Philosophy*. 82(11). pp. 607 – 616.

Hempel, C., 1980. Comments on Goodman's way of Worldmaking. *Synthese.* 45(2). pp. 193 – 199.

Husserl, E., 1931/1960. *Cartesian Meditations*. Translated from French by D. Cairns. The Hague, NL: Martinus Nijhoff Publishing.

International Committee of the Red Cross, 1977. *Additional Protocol 1: Article 52*. [online] Available at: <https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Article.xsp?action=openDocument&documentId=F08A9BC78AE360B3C12563CD0051DCD4>. Accessed 21 April 2021.

Jakobson, R., 1960. Closing Statement: Linguistics and Poetics. In: *Style in Language.* T. A. Sebeok (ed). Cambridge, MA: MIT Press. pp. 350 – 377.

Jiang, F., Jiang, Y., Zhi, H., et al., 2017. Artificial Intelligence in Healthcare: Past, Present, and Future. *Stroke and Vascular Neurology*. 2. pp. 230 – 242.

Johnson, D. G., 2006. Computer Systems: Moral Entities but not Moral Agents. *Ethics and Information Technology*. 8. pp. 195 – 204.

Jonas, H., 1979/1985. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Translated from German by H. Jonas and D. Herr. Chicago, IL: Chicago University Press.

Kane, R., 2005. *A Contemporary Introduction to Free Will*. Oxford, UK: Oxford University Press.

Kant, I., 1787/2007. *Critique of Pure Reason*. Translated from German by M. Weigelt. London, UK: Penguin Classics.

Kaplan, J., 2016., *Artificial Intelligence: What Everyone Needs to Know*. Oxford, UK: Oxford University Press.

Kenny, A., 2007. *Ancient Philosophy: Vol. 1*. Oxford, UK: Oxford University Press

Keynes, J. M., 1930. Economic Possibilities for our Grandchildren. In: *The Essential Keynes*. 2015. R. Skidelsky (ed). London, UK: Penguin Classics. pp. 75 – 86.

Kilkenny, M. F., and Robinson, K. M., 2018. Data Quality: "Garbage in – Garbage Out". *Health Information Management Journal*. 47(3). pp. 103 – 105.

Kim, J., 2005., *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.

Kim, T. W., and Scheller-Wolf, A., 2019. Technological Unemployment, Meaning in Life, Purpose of Business and the Future of Stakeholders. *Journal of Business Ethics*. 160. pp. 319 – 337.

Kirk, R., 1994. *Raw Feeling.* Oxford, UK: Clarendon Press.

Kuhn, T. S., 1977. *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago, IL: University of Chicago Press.

Laudan, L., 1981. A Confutation of Convergent Realism. *Philosophy of Science*. 48(1). pp. 19 – 49.

Leibniz, G.W., 1765/1896. *New Essays Concerning Human Understanding*. Translated from French, German and Latin by A.G. Langley. London: MacMillan & Co., Ltd.

Lenin, V. I., 1894/1977. What the "Friends of the People" are and How they Fight the Social-Democrats. In: *Collected Works, Volume 1*. Translated from Russian. Moscow, RU: Progress Publishers. pp. 129 – 332.

Lenin, V. I., 1896/1972. Frederick Engels. In: *Collected Works, Volume 2.* Translated from Russian by G. Hanna. Moscow, RU: Progress Publishers. pp. 15 – 28.

Lenin, V. I., 1908/1977. Materialism and Empirio-criticism. In: *Collected Works, Volume 14.* Translated from Russian by A. Fineberg. Moscow, RU: Progress Publishers.

Lenin, V. I., 1916/1974. The Junius Pamphlet. In: *Collected Works, Volume 22.* Translated from Russian by Y. Sdobnikov. Moscow, RU: Progress Publishers. pp. 305 – 319.

Lenin, V. I., 1917/1974. The State and Revolution. In: *Collected Works, Volume 25.* Translated from Russian. Moscow, RU: Progress Publishers. pp. 385 – 498.

Lepri, B., Staiano, J., Sangokoya, D., et al., 2017. The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good. In: *Transparent Data Mining for Big and Small Data*. T. Cerquitelli, D. Quercia and F. Pasquale (eds). Berlin, DE: Springer. pp. 3 – 24.

Levine, N., 1975. *The Tragic Deception: Marx Contra Engels*. Oxford, UK: Clio Books.

Liang, F., Das, V., Kostyuk, N., and Hussain, M. M., 2018. Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure. *Policy & Internet*. 10(4) pp. 415 – 453.

Libet, B., 1985. Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioural and Brain Sciences*. 8. pp. 529 – 566.

Lin, P., Bekey, G., and Abney, K., 2008. *Autonomous Military Robotics: Risk, Ethics, and Design.* [pdf] Available at: <http://ethics.calpoly.edu/onr_report.pdf>. Accessed 12 April 2021.

Locke, J., 1689/1997. *An Essay Concerning Human Understanding*. London, UK: Penguin Classics.

Marcuse, H., 1964/2004. *One-Dimensional Man*. 2nd Ed. Oxford, UK: Routledge.

Marcuse, H., 1965. Industrialization and Capitalism in the Work of Max Weber. Translated from German by J. J. Shapiro. In: *Negations: Essays in Critical Theory*. 2009. London, UK: Mayfly.

Marcuse, H., 1977. Heidegger's Politics: An Interview. In: *Heideggerian Marxism*. 2005. R. Wolin and J. Abromeit (eds). Lincoln, NB: University of Nebraska Press

Markkula, G., 2015. 'Answering Questions about Consciousness by Modelling Perception as Covert Behavior' *Frontiers in Psychology*. 6(803). [online] Available at: <https://doi.org/10.3389/fpsyg.2015.00803>.

Marx, K., 1844/1992. Economic and Philosophical Manuscripts. In: *Early Writings*. Translated from German by R. Livingstone and G. Benton. London, UK: Penguin Classics. pp. 279 – 400.

Marx, K., 1847/1973. *The Poverty of Philosophy*. Moscow, RU: Progress Publishers.

Marx, K., 1859/1977. *A Contribution to the Critique of Political Economy*. Translated from German by S. W. Ryazanskaya. Moscow, RU: Progress Publishers.

Marx, K., 1867/2008. *Capital: An Abridged Edition*. D. McLellan (ed). Oxford, UK: Oxford University Press.

Marx, K., 1873/2008. Afterword to the Second German Edition. In: *Capital: An Abridged Edition*. D. McLellan (ed). Oxford, UK: Oxford University Press.

Matthias, A., 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*. 6. pp. 175 – 183.

McAfee, A., and Brynjolfsson, E. 2016. Human Work in the Robotic Future: Policy for the Age of Automation. *Foreign Affairs.* 95(4) pp. 139 – 150.

McAfee, A., and Brynjolfsson, E., 2017. *Machine Platform Crowd: Harnessing Our Digital Future*. New York, NY: W. W. Norton and Company.

McClelland, T., 2017. Against Virtual Selves. *Erkenntnis*. 84(1). pp. 21 – 40.

Mele, A. R., 2014. *Free: Why Sciences Hasn't Disproved Free Will*. Oxford, UK: Oxford University Press.

Melnyk, A., 2003. *A Physicalist Manifesto: Thoroughly Modern Materialism*. Cambridge, UK: Cambridge University Press.

Merleau-Ponty, M., 1945/2000. *Phenomenology of Perception*. Translated from French by C. Smith. London, UK: Routledge.

Metzinger, T., 2004. *Being No One: The Self-Model Theory of Subjectivity.* Cambridge, MA: Bradford, MIT Press.

Metzinger, T., 2009. *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. New York, NY: Basic Books.

Mulhall, S., 2005. *Heidegger and Being and Time*. Oxford, UK: Routledge.

Nagel, T., 1972. War and Massacre. *Philosophy and Public Affairs*. 1. pp. 123 – 144.

Office for National Statistics, 2017. *UK Government Expenditure on Science, Engineering and Technology*. Newport: Office for National Statistics. [pdf] Available at: <https://www.ons.gov.uk/economy/governmentpublicsectorandtaxes/researchanddevelopmentexpenditure/bulletins/ukgovernmentexpenditureonscienceengineeringandtechnology/2015/pdf>. Accessed 20 November 2017.

Olson, E. T., 2007. *What Are We? A Study in Personal Ontology*. Oxford, UK: Oxford University Press.

Papineau, D., 1993. *Philosophical Naturalism*. Oxford, UK: Blackwell.

Parfit, D., 2012. We Are Not Human Beings. *Philosophy*. 87(1). pp. 5 – 28.

Pattison, G., 2000. *The Later Heidegger*. Oxford, UK: Routledge.

Piketty, T., 2014. *Capital in the Twenty-First Century*. Translated from French by A. Goldhammer. Cambridge, MA: Belknap/Harvard.

Pinker, S., 2018. *Enlightenment Now: The Case for Reason, Science, Humanism and Progress.* London, UK: Allen Lane.

Place, U. T., 1956. 'Is Consciousness a Brain Process?' *British Journal of Psychology*. 47(1). pp. 44 – 50.

Plato., n.d. Phaedo. In: *The Last Days of Socrates*. 2003. London, UK: Penguin Classics. pp. 97 – 200.

Polkinghorne, J., 2002. *Quantum Theory: A Very Short Introduction*. Oxford, UK: Oxford University Press.

Putnam, H., 1967. The Nature of Mental States. In: N. Block, ed. 1980. *Readings in Philosophy of Psychology: Volume One*. London, UK: Methuen. pp. 223 – 232.

Quine, W. V. O., 1951. Two Dogmas of Empiricism. *Philosophical Review*. 60(1). pp. 20 – 43.

Rockström, J., Steffen, W., Noone, K., et al., 2009. A Safe Operating Space for Humanity. *Nature*. 461. pp. 472 – 475.

Roff, H. M., 2014. The Strategic Robot Problem: Lethal Autonomous Weapons in War. *Journal of Military Ethics*. 13(3). pp. 211 – 227.

Rorty, R., 1979/2009. *Philosophy and the Mirror of Nature: 30th Anniversary Ed*. Princeton, NJ: Princeton University Press.

Rorty, R., 1991. *Objectivity, Relativism and Truth*. Cambridge, UK: Cambridge University Press.

Russell, S. J., and Norvig, P., 2010. *Artificial Intelligence: A Modern Approach*. 3rd Ed. Upper Saddle River, NJ: Prentice Hall.

Santoni de Sio, F., and Mecacci, G., 2021. For Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*. [online] Available at: <https://doi.org/10.1007/s13347-021-00450-x>.

Sartre, J-P., 1938/2000. *Nausea*. Translated from French by R. Baldick. London, UK: Penguin Modern Classics.

Scharre, P., and Horowitz, M. C., 2015. An Introduction to Autonomy in Weapons Systems. *CNAS*. [online] Available at: <http://files.cnas.org/documents/Ethical-Autonomy-Working-Paper_021015_v02.pdf>. Accessed 7 April 2021.

Scharre, P., 2018. *Army of None: Autonomous Weapons and the Future of War*. New York, NY: W. W. Norton & Company.

Schwab, K., 2016. *The Fourth Industrial Revolution*. London, UK: Penguin.

Searle, J. R., 1980. Minds, Brains and Programs. *The Behavioural and Brain Sciences*. 3. pp. 417 – 457.

Searle, J. R., 1992. *The Rediscovery of the Mind.* Cambridge, MA: MIT Press.

Sellars, W., 1997. *Empiricism and the Philosophy of Mind*. Cambridge, MA: Harvard University Press.

Smart, J. J. C., 1959. 'Sensations and Brain Processes' *Philosophical Review*. 68(2). pp. 141 – 156.

Smith, A., 1776/1976. *An Inquiry into the Nature and Causes of the Wealth of Nations.* E. Cannan (ed). Chicago, IL: University of Chicago Press.

Snowdon, P.F., 1989. On Formulating Materialism and Dualism. In: *Cause, Mind and Reality: Essays Honoring C. B. Martin*. 1989. J. Heil (ed). pp. 137 – 158.

Sornette, D., and von der Becke, S., 2011. Crashes and High Frequency Trading: An Evaluation of Risks Posed by High-Speed algorithmic Trading. [pdf] London, UK: Government Office for Science. Available at: <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/289016/11-1226-dr7-crashes-and-high-frequency-trading.pdf>. Accessed 22 March 2021.

Sparrow, R., 2016. Robots and Respect: Assessing the Case Against Autonomous Weapons Systems. *Ethics and International Affairs*. 30(1). pp. 93 – 116.

Stalin, J., 1938/1969. *Dialectical and Historical Materialism*. New York, NY: International Publishers.

Stannard, R., 2008. *Relativity: A Very Short Introduction*. Oxford, UK: Oxford University Press.

Stetson, D. S., Albers, J. W., Silverstein, B. A., and Wolfe, R. A.,1992. Effects of Age, Sex, and Anthropometric Factors on Nerve Conduction Measures. *Muscle & Nerve*. 15. pp. 1095 – 1104.

Steward, H., 2012. *A Metaphysics for Freedom*. Oxford, UK: Oxford University Press.

Steward, H., 2020. Agency as a Two-Way Power: A Defence. *The Monist*. 103. pp. 342 – 355.

Stoljar, D., 2010. *Physicalism*. London, UK: Routledge.

Strawson, G., 1986. *Freedom and Belief*. Oxford, Oxford University Press.

Svetlova, E., 2012. On the performative power of financial models. *Economy and Society*. 41(3). pp. 418 – 434.

Tallis, R., 2011/2016. *Aping Mankind*. London, UK: Routledge Classics.

Tallis, R., 2020. *Seeing Ourselves: Reclaiming Humanity From God and Science.* Newcastle-upon-Tyne, UK: Agenda Publishing.

Topal, E. J., 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*. 25. pp. 44 – 56.

Turing, A. M., 1950. Computing Machinery and Intelligence. *Mind*. 59(236). pp. 433 – 460.

Umbrello, S., Torres, P., and De Bellis, A. F., 2020. The Future of War: Could Lethal Autonomous Weapons Make Conflict More Ethical? *AI & Society.* 35. pp. 273 – 282.

Vallor, S., 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford, UK: Oxford University Press.

Van Inwagen, P., 1983. *An Essay On Free Will*. Oxford, UK: Clarendon Press.

Van Inwagen, P., 1990. *Material Beings*. Ithaca, NY: Cornell University Press.

Veltman, A., 2016. *Meaningful Work.* Oxford, UK: Oxford University Press.

Walsh, T., 2018. Expert and Non-expert Opinion About Technological Unemployment. *International Journal of Automation and Computing*. 15(5) pp. 637 – 642.

Wellman, M. P., and Rajan, U., 2017. Ethical Issues for Autonomous Trading Agents. *Minds and Machines*. 27. pp. 609 – 624.

Wolf, S., 2012. *Meaning in Life and Why it Matters*. Princeton, NJ: Princeton University Press.

Zahavi. D., 2016. The end of what? Phenomenology vs. speculative realism. *International Journal of Philosophical Studies*. 24(3) pp. 289 – 309.