CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Artificial intelligence as religion: an evolutionary account and philosophical study
Authors	Darby, Max Hollis
Publication date	2020-10-01
Original Citation	Darby, M. H. 2020. Artificial intelligence as religion: an evolutionary account and philosophical study. MPhil Thesis, University College Cork.
Type of publication	Masters thesis (Research)
Rights	© 2020, Max Hollis Darby. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-10-30 04:29:50
Item downloaded from	https://hdl.handle.net/10468/12384



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Artificial Intelligence as Religion: An Evolutionary Account and Philosophical Study.



Author: Max Hollis Darby.

Qualification: Research Masters (MPhil).

Institution: University College Cork.

School/Department: College of Arts, Celtic Studies and Social Sciences/Philosophy.

Head of Dept.: Prof. Don Ross.

Supervisors: Dr. Órla Murphy and Dr. Joel Walmsley.

Year of Submission: 2020

Table of Contents

DECLARATION.....5

ABSTRACT.....6

1 SECTION 1: INTRODUCTION AND LITERATURE REVIEW. 7

1.1 INTRODUCTION.....7

1.2 RELIGION:.....8

1.2.1 DO RELIGIONS REQUIRE BELIEF IN A SUPERNATURAL ENTITY?.....8

1.2.2 RELIGION AS A FUNCTION IN SOCIETY. 10

1.2.3 RELIGION AS DISTINCT FROM SPIRITUALITY. 12

1.2.4 DEFINITION OF RELIGION 13

1.3 EVOLUTION: 16

1.4 ARTIFICIAL INTELLIGENCE, TECHNO-SOCIAL ENVIRONMENTS, AND TAKE-OFFS. 20

1.4.1 GENERAL AI. 20

1.4.2 NARROW AI 22

1.4.3 AI ENVIRONMENTS AS DISTRIBUTED SYSTEMS: 23

1.4.4 SEPARATING THE SYSTEM FROM THE OWNER. 25

1.4.5 AI ENVIRONMENTS EMBODYING A VALUE-SYSTEM 27

1.4.6 AI ENVIRONMENTS AS AN UNINTENDED BY-PRODUCT. 27

1.4.7 TAKE-OFF SCENARIOS. 29

1.4.8 SUMMARY OF THE TECHNOLOGICAL CONTEXT OF THIS RESEARCH: 31

1.5 PHILOSOPHY OF MIND 32

1.5.1 THE MIND AS A RESULT OF, BUT NOT LIMITED TO, PROCESSES IN THE BRAIN. 33

1.5.2 THE MIND RETAINS EXECUTIVE CONTROL AND THE ABILITY TO DO OTHERWISE. 34

1.5.3 SELF-DETERMINATION AND MORAL RESPONSIBILITY AS SUFFICIENT FOR INDIVIDUAL FREE-WILL. 35

1.5.4 AN INDIVIDUAL THAT CAN DETERMINE THEIR OWN VALUES AND BE MORALLY RESPONSIBLE FOR THEM IS FREE. 37
1.5.5 IS AN OUTSOURCED DECISION AN EXTENSION OF THE MIND? 38
1.6 CONCLUSION 40
1.6.1 RELIGION. 40
1.6.2 EVOLUTION. 41
1.6.3 ARTIFICIAL INTELLIGENCE. 41
1.6.4 PHILOSOPHY OF MIND. 42
1.7 CONCLUSION OF LITERATURE REVIEW AND INTRODUCTION TO THE RESEARCH 44

2 ORIGINS OF RELIGIONS AND AI ENVIRONMENTS. 46

2.1 INTRODUCTION AND APPROACH. 46
2.2 FUNDAMENTAL ORIGIN OF RELIGION: UNCONSCIOUS THOUGHT BECOMES MYTH. 46
2.3 FUNDAMENTAL ORIGINS OF AI: RAW DATA BECOMES PROFILING. 48
2.4 ARE BIG DATA PROFILES ACCURATE REPRESENTATIONS? 51
2.5 CONCLUSION 53

3 EVOLUTION OF RELIGIOUS CONCEPTS. 55

3.1 INTRODUCTION. 55
3.2 MYTH AS A SOCIETAL FUNCTION. 55
3.3 INTRODUCING INTERMEDIARIES : BIOLOGICAL ROOTS. 56
3.4 INTRODUCING INTERMEDIARIES: BIOLOGY AND THE ROLE OF RITUAL. 59
3.4.1 LEVERAGING BIOLOGICAL EVOLUTION 59
3.4.2 RITUALISED BEHAVIOUR 61
3.4.3 EVALUATION OF RITUALISED BEHAVIOUR IN AI ENVIRONMENTS 63
3.5 RELIGIOUS EVOLUTION: RELIGIONS AS MEMES. 68
3.5.1 PSYCHOLOGICALLY PRIMED: HYPERACTIVE AGENT DETECTION DEVICE (HADD): 71

3.5.2 PSYCHOLOGICALLY PRIMED: ARE ALGORITHMS THE NEW SUPERNATURAL? 73
3.5.3 PSYCHOLOGICALLY PRIMED: SUPERNATURAL AS OPAQUE COMPLEXITY 75
3.5.4 PSYCHOLOGICALLY PRIMED: PERSONAL SACRIFICE (EXCHANGES OF VALUE) 77
3.6 FINAL EVALUATION OF AI ENVIRONMENT CONCEPTS. 80
3.6.1 MAKING RECALL AND COMMUNICATION EASY: 80
3.6.2 TRIGGERING EMOTIONAL PROGRAMS: 81
3.6.3 CONNECTING TO OUR SOCIAL MIND: 82
3.6.4 BECOMING PLAUSIBLE AND DIRECT BEHAVIOUR: 84
3.7 CONCLUSION 84

4 SPECULATION ABOUT THE FUTURE. 86

4.1 REDUCED DIMENSIONS OF EXPERIENCE 86
4.2 DEVELOPING AN INABILITY TO SELF-DETERMINE VALUES. 89
4.3 BECOMING STIMULUS-RESPONSE MACHINES & THE LOSS OF INDIVIDUALITY. 92
4.4 AN ENVIRONMENT FOR INDIVIDUALITY 95
4.5 CAN AN AI ENVIRONMENT QUALIFY AS AN EXTENSION TO THE MIND OF THE INDIVIDUAL? 97
4.6 WHERE SHOULD WE BE FOCUSING OUR EFFORTS? 102

BIBLIOGRAPHY 105

Declaration.

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

Abstract.

Religions and religious behaviours have been documented in biological and evolutionary terms. This research considers how religions emerged as distributed, de-centralised biological extensions and evolved into centralised cultural organisations. This provides a model of the evolutionary mechanisms that contributed to the origin, development, and proliferation of religions. It establishes that religions encouraged, curated, and leveraged a specific mentality that has not disappeared despite humanity's move toward secularism. This research interrogates whether the religiously primed mind will attempt to fill a cognitive void with artificial intelligence (AI) systems in a post-religious society. This comparison provides an evolutionary account for how AI systems will use existing religious mechanisms and behavioural tendencies to develop and proliferate from de-centralised extensions of cognition to centralised cultural systems.

This research finds that the scenario described above has significant implications with regard to human individuality, moral responsibility, and individual freedom. The thesis will conclude with a proposal for the necessary requirements for retaining these three features in a future where significant amounts of cognitive processes are outsourced to AI systems.

1 Section 1: Introduction and Literature Review.

1.1 Introduction

This thesis will synthesize concepts and patterns from religious studies, theories of evolution, information technology, and Philosophy of Mind. In this respect, the analysis of the available literature is segmented into the categories of each subject.

The subjects and questions that require analysis are:

- Religion: How is this defined? What are the necessary conditions needed to categorise something as a religion? How can we recognise a religion?
- Biological and cultural evolution: What are these? How does the concept of evolution work? How do biological and cultural evolution apply to religion?
- Information technology: How is artificial intelligence defined within this research? What are techno-social environments? What are the potential take-off patterns?
- Philosophy of Mind: Which theories examine the implications of outsourcing cognition?

These questions will be addressed in this section, with the intent that the answers can support a comparison between the emergence of artificial intelligence (AI) and religion, leading to an evaluation of AI *as* a religion, which will warrant a further exploration into the implications for human cognition.

1.2 Religion:

The purpose of the texts documented within this subsection is to understand the difference between a religion and a sense-making tool (or value-system). All religions are sense-making tools, but not all sense-making tools are religions.

At a high-level summary, and for the purposes of my discussion, a religion will be considered as: *any concept- or collection of concepts- that prescribes a value system that shapes human behaviour and that also:*

1) *Bridges a gap in explicability and irreducible complexity through requiring belief in an entity whose abilities are incomprehensible to the human mind.*
2) *Encourages a reluctance to further critical inquiry which is justified by a predetermined representation of an absolute order of the universe.*

This definition may include concepts that are commonly considered religious, such as Christianity or Judaism, but does not include value-systems such as capitalism or communism, as while these concepts are bound to a fixed representation of absolute order (i.e. trust in the free market), they do not conceal their complexity behind something irreducible or inexplicable. Examples of each will documented throughout this discussion.

1.2.1 Do religions require belief in a supernatural entity?

Daniel Dennett addresses an issue in defining religion in the introductory pages of his book *Breaking The Spell* (2007). He raises the issue that too broad of a definition can encapsulate too many concepts, such as the theory of evolutionary biology, which he notes could have implications for 'Legal protection, honor, prestige, and a *traditional exemption from certain sorts of analysis and criticism*' (Dennett 2007, 9). Dennett's definition of religion thus narrows itself to a 'social system whose participants avow belief in a supernatural agent or agents whose approval is to be sought' (9). This definition tends more towards traditional conceptions of religions such as Christianity, Judaism and Islam, which all have supernatural agents that play significant roles in their doctrine. Notably this

definition could exclude religions such as Buddhism which do not have any formal supernatural gods. Curiously, 'whose approval is to be sought' implies that there is a list of values for which a follower can be evaluated against, which I interpret as Dennett also prescribing that a religion must have some sort of 'representation of absolute order' (i.e. a definition of what is good or bad) that must also be available to the follower. This is to say, that the follower must be able to identify a value system prescribed by the religion, which they can live in accordance with, in order to seek approval. In a roundabout way, this could also include Buddhism in Dennett's definition, because it does prescribe a way of living which is open to evaluation. I take Dennett's definition to include approval in the absence of an approver. A follower can still act in a way that they think Buddha *would have approved of, if he was still alive*.

However, Dennett's definition specifically necessitates the belief in a supernatural agent, which may be too constraining for the re-use of this definition in other contexts. His book was written at a pivotal point in time, when critical analysis of religion was only beginning to be culturally acceptable. His definition may have been formulated in a way as to include the major religions that would have posed most opposition to his analysis. Spokespeople for these religions would have used any broadness of scope to devalue his claims in rebuttal (as is addressed immediately prior to the definition where he worries that too broad of a definition could include biological evolution).

It is more appropriate to understand Dennett's definition instead by what he proposes as the core phenomena of religion. 'The core phenomena of religion *invokes gods* who are effective agents in real time, and *who play a central role in the way the participants think about what they ought to do* (italics my own)' (11). In this way, we can evaluate religions with regard to how their representations of gods shape human action, rather than what features they may have. For example, take someone practicing Christianity. They take part in the rituals, masses, and prayers. This person confesses sins and contemplates how their actions affect the world around them. However, this person does not believe in God. This is not an incompatible combination, and the lack of belief does not make their actions disingenuous. The question this poses is: is this person *religious*? Evolutionary biologist Bret Weinstein (who also informs the evolutionary portion of this review) believes that religions can still exist as useful tools, and that we should not 'throw out the baby with the bathwater' (*Richard Dawkins & Bret Weinstein - Evolution* n.d.) , so to speak. Informed by

this interpretation of religion as the function it plays, we can say that a person practicing a religion but not believing in the supernatural god is using the religion *as a sense-making tool*. They are using the religious practices to enhance their life, perhaps to help them make sense of ethical doubts they have or to reconcile differences, or perhaps just for the comfort of tradition. This distinction marks a significant necessity of what should be present in order to declare something a religion. Dennett describes this as a 'supernatural agent', however this will have to be broadened if the definition is to be re-used. Whatever it is called, this necessary agent must be *believed* in.

1.2.2 Religion as a function in society.

Understanding religions through the lens of how they shape human behaviour is adopted by Yuval Noah Harari in his book *Homo Deus* (2016). Considering that Harari will not likely come up against opposition by religious commentators for his use of definitions (his book was written over a decade later than Dennett's, and the subject matter was not exclusively aimed at debunking religions), his definition of religion is decisively more liberal than Dennett's and is more focused on how religions shape how humans understand the world and how religions direct human behaviour.

Harari does not provide a one-line definition of religion however he does allude to the phenomena a religion produces in the second part of his book: 'Homo Sapiens Gives Meaning to the World'. He chooses to define religion by its function in society, instead of by the features it has such as 'belief in gods', or 'faith in supernatural powers' (Harari 2016, 211). This can be summarised by his statement that 'religion asserts that we humans are subject to a system of moral laws that we did not invent and that we cannot change'.

He bases his definition of religion from a broad human-centric point of view, which considers religions as social constructs that 'organise mass cooperation' through a 'common network of stories' (170).

Harari draws a comparison between these religions, or collection of gods, to modern brands in the sense that they are 'fictional legal entities that own property, lend money, hire employees and initiate economic enterprises'. Using this analogy, Harari helps the reader understand religion as a corporation, which organises humans under a common

system of values, or as he describes it: ‘a well-defined contract with predetermined goals’ (214). Here we can see a strong correlation with Dennett’s ‘core phenomena’, where religions ‘play a central role in the way participants think about what they ought to do’, and Harari’s definition: ‘The very clarity of this deal allows society to define common norms and values that regulate human behaviour’ (215).

However, Harari’s definition differs from Dennett’s in the sense that Harari’s includes ‘common network(s) of stories’ such as Buddhism and Daoism (neither of which have supernatural deities) and communism, Nazism, and liberalism (none of which are generally considered religions). To Harari, what makes these concepts religions is that they ‘argue that these so-called superhuman laws are natural laws, and not the creation of this or that god’ (212).

Harari’s view of religion as a social construct that ‘organises mass cooperation’ is substantiated by a recent piece of research that performed a cross-cultural comparative analysis on records from ‘414 societies that span the past 10,000 years from 30 regions around the world’ (Whitehouse et al. 2019). This study analyses the association between moralizing gods and social complexity. Moralizing gods are defined as the supernatural deities of prosocial religions that ‘punish moral violations in interactions between humans [and] … human moral transgressions’ (226). This analysis confirmed the association between the two, but also ‘reveal[ed] that moralizing gods follow–rather than precede–large increases in social complexity’ (226). Further analysis indicates that ‘moralizing gods are not a prerequisite for the evolution of social complexity, but they may help to sustain and expand complex multi-ethnic empires’ (226).

While this study suggests that moralizing gods do not cause (precede) the rise of a complex society, it posits that ‘they may represent a cultural adaptation that is necessary to maintain cooperation in such societies once they have exceeded a certain size’ (228). This conclusion is drawn from the observation that only one society out of ten that did not develop moralizing gods became a complex society (the Inca Empire).The aspect of this study that substantiates Harari’s view is that it also suggests that ritual practices played a critical role in ‘the initial rise of social complexity’, even outside of their role within the religion.

In summary, this study suggests that Harari is correct to define religion as a 'deal that allows society to define common norms and values that regulate human behaviour', but that Dennett is also correct in his claim that there must–eventually–be a supernatural agent whose approval must be sought (if the society is to remain stable, that is). This is to say: a religion, if it is to have longevity, must eventually prescribe a *punishable* system of values.

Harari's overall definition of religion is too liberal to be used directly in this thesis, however the discussion is informed by his approach to understanding a religion as a social contract within a society. In this way, we can recognise that an identifying factor of a religion is that it prescribes a common value-system and influences behaviour. Whitehouse et al. informed us that although it is typically true that the prescribed value-system is a *punishable* set of values for successful religions, it is not necessary that this be the case in the process of actually identifying *any* religion, and will therefore not be included as a necessary component of the definition. However, this test will inform a later discussion on effective religious features.

1.2.3 Religion as distinct from Spirituality.

It is at this point important to identify any distinctions between religion and spirituality, for fear that it may cloud the definition. Harari touches upon this in *Homo Deus*, where he claims that religion and spirituality are different in the sense that 'religion is a deal, whereas spirituality is a journey' (214). By this, Harari means that religion will give you a value system to follow, without an explanation or internal realisation that the value is correct. Spirituality, he contrasts, 'begins with some big question, such as who am I? What is the meaning of life? What is good?' (215).

Alan Watts conveys a similar distinction in his book *The Wisdom of Insecurity*, in which he interprets belief and faith separately. He describes belief (religion in Harari's terms) as 'an insistence that the truth is what one would … wish it to be', i.e. clinging to a prescribed ideal. Watts contrasts this definition of belief with a definition of faith (spirituality to Harari) as an 'unreserved opening of the mind to the truth, whatever it may turn out to be', i.e. letting go (p.23). Summarised in his words as: 'Belief clings, but faith lets go'. This has

a direct correlation with Harari's distinction of religion as a fixed 'deal', and spirituality as a 'journey', where a participant can cling to a deal, or let go to be taken on a journey.

This is not to say that religion might not be a way to spiritual attainment, but it does draw a distinction and suggests that they are not the same thing. This implies that spiritual attainment may be accessible by other methods outside of religion. By way of analogy: religion may be a boat which can take you down the river of spirituality, but they should not be confused for each other, and many other types of boats exist.

In his book *Modern Man in Search of a Soul* (2001), Carl Jung highlights four parts of life that are only accessible through what he defines as 'the clergyman': 'faith, hope, love, insight'. He states that these 'gifts of grace' can 'neither… be taught nor learned, neither given nor taken, neither withheld nor earned' but are a result of experience (231).
Jung's use of the term 'clergymen' implies that these four experiences are only obtainable through religion, however I posit that these essential parts of life are not actually only accessible through religion, but are achievable through spirituality, and that Jung has conflated the boat for the river. For this reason, I have intentionally omitted any types of spirituality from my definition of religion, because the literature suggests that they are fundamentally distinct phenomena.

1.2.4 Definition of Religion

Dennett's and Harari's definitions of religion have highlighted the two distant points on a hypothetical scale of religious definitions. Dennett's definition is concerned with how one could identify a religion as they have manifested up until now. Harari's definition allows the categorisation more or less at the inception of any set of shared values, but does so at the cost of being able to identify the concepts that differentiate religions from sense-making tools. Any definition more liberal that Harari's risks being too broad to remain within the scope of being able to examine anything substantial, and any definition more conservative than Dennett's risks limiting any comparisons to those only between established religions. Both definitions are appropriate for the contexts that they exist in, but neither can be directly used in this instance.

The definition of religion for the purpose of this research must sit somewhere on this scale, to mark the line where a concept moves from being defined as a religion toward being a tool. Dennett's definition indicates that it is a supernatural agent, but a supernatural agent appears to be an embodiment of the concept, rather than the actual differentiating feature in itself. The 'death of God' acts as a proxy for the death of religion because God is the embodiment of religion. Whereas the death of Marx did not act as a proxy for the death of Communism.

This issue with defining what constitutes 'embodiment' is best highlighted with an example central to this research: Artificial Intelligence is just a collection of algorithms that are used as a tool to make sense of data. Some people appreciate that the automated decision-making apparatus consists of mathematics and data-feeds. This is comparable to the free-market. Trustees of this system may not fully understand it, but there is a trust that, in its complexity, it is more intelligent than the trustee, but also an inherent transparency that it is not irreducibly complex. On the other side, there are people who refer to AI systems as 'the algorithms' and wrongly attribute a sense of incomprehensibility and omniscience to something that is, in reality, neither.

This is the differentiating feature that must be identified, but not present in either definition (although alluded to in Dennett's). To summarise, we can consider this feature as; a perceived gap in explicability and irreducible complexity, that is compensated for through belief in an entity whose abilities are incomprehensible to the human mind.

This research is ultimately concerned with how a religion begins, which occurs before it would be classed as a religion under Dennett's view. It is also concerned with how a concept that is not commonly considered a religion can actually be classified as one, as in Harari's point surrounding communism and Nazism. Finally, it is concerned with how a concept influences mass human behaviour, which is central to Harari's definition. However, it does draw a defining distinction between religions and tools.

For these reasons, and for the purpose of this research, we will consider a religion as:

- *any concept- or collection of concepts*: Myths, laws, doctrines, algorithms.

- *that prescribes a value-system that shapes human behaviour*: Prescribes action (or inaction) based on a judgement of value.
- *Bridges a gap in explicability and irreducible complexity through requiring belief in an entity whose abilities are incomprehensible to the human mind*: "the Lord works in mysterious ways…", "the algorithms said…".
- *and encourages a reluctance to further critical inquiry justified by a representation of an absolute order of the universe*: Determinism, Computational reductionism (Datafication), Creationism, The Way/Tao.

It is important to note that this list of criteria purposely does not include scientific theories, due to the fact that while they formulate a fixed representation of absolute order, they do not actually prescribe a value-system. In areas where scientific-based value systems are prescribed, it is from a human interpretation of an objective quantity (e.g. Sam Harris' *Moral Landscape* (2012), where value can be derived from neuroscientific measurements, to be discussed later).

As previously stated, it is important for this research to understand how religions were formed. They did not manifest themselves in the world as the powerful, centralised, precise organisations we see today. As Harari's definition permits us to investigate, religions–like most organisations–had humble beginnings in the minds of (literally) just a couple of humans, however we will be careful to not confuse sense-making tools with religions.

1.3 Evolution:

The ability to examine religion through the framework of biological and cultural evolutionary theories has facilitated research into how religions formed and propagated across generations and cultures. Prior to this, religions were able to defend themselves against inquiry with claims of cosmically endowed virtue. This appeal became more difficult to defend with the emergence of scientific and empirical research (which, ironically, was first encouraged by the church in order to explore God's glorious creation).

Biological evolution is concerned with how biological organisms change across time, how imperceptible changes between each generation of species can result in drastically different entities, given enough generations (or 'cycles' (Dennett 2014, 255)). This process provides an empirically satisfactory account for the origins and variance of organisms on the planet, without the need to resort to any intelligent grand-designer (as religions do). While the theory of evolution predates the actual discovery of the mechanisms that account for this process (such as DNA and genes), it is now understood as the study of how genes mutate and replicate, based on their robustness in the surrounding environment, which can provide a satisfactory account for the biological arrival at even the most complex organs such as the eyeball (which are 'sometimes erroneously described as 'irreducibly complex'') (Dawkins 2016, 148).

Cultural evolution is similarly concerned with incremental changes across cycles, however not of biological life, but of ideas (or 'memes' (Dawkins 2016, 222; Dennett 2007, 341)). In the same way that theories of biological evolution can account for changes in the gene-pool, theories of memetic evolution can account for changes in what Richard Dawkins refers to as 'the memeplex' (2016, 228).

Much to the detriment of religious thinking, evolutionary theories have been able to explain the emergence of complex life without the need for any all-mighty designer. Daniel Dennett calls this 'competence without comprehension' in his book, *Intuition Pumps and Other Tools for Thinking* (2014, 105). Likewise, if religions themselves are memes, and memes can be accounted for using evolutionary theories, then religions themselves (not just their claims) are subject to an evolutionary account.

Anthropologist Pascal Boyer, philosopher Daniel Dennett, and evolutionary biologist Richard Dawkins all use evolutionary theories in order to explain how religions could have emerged, and how they evolved across cultures. They use these empirical methods in order to explain 'religion as a natural phenomenon' (the subtitle of Dennett's book (2007)) and to uncover the foundational mechanisms that created such complex systems, as a way of demonstrating an alternative to the theory of cosmically endowed truth that religions have used to verify their legitimacy.

A benefit of using scientific theories to explain these types of phenomena is that they can also be used predictively. Bret Weinstein, in a debate with Richard Dawkins, declared his admiration for this in the work of evolutionary biologist George Williams and his theories of senescence (*Richard Dawkins & Bret Weinstein - Evolution* n.d.). He praised the fact that Williams' paper openly predicted that if his theories were correct, that particular patterns would be observable in nature (which was, in fact, the case).

The universality of these theories means that the aspects of biological and cultural evolution can be used to make predictions about any phenomena that contain the same apparatus. As far as I have been able to identify from the key texts, researchers have only used evolutionary theories to explain existing religious phenomena, however Harvey Whitehouse and colleagues do illustrate how these explanations can be formulated to create models of societies, which can be used for making predictions (e.g. they predict that any society devoid of a punishable system of values or a moralising god, will collapse) (2019) . Given the comparison that this research makes between religions and artificial intelligence, it is appropriate that these evolutionary theories be used to make a model to predict how AI environments will develop.

It is important to note that biological and cultural evolution are used to explain separate phenomena. Some evolutionary theorists argue that cultural evolution can (and should) be reduced to biological evolution because cultural evolution is ultimately a product of biology. Bret Weinstein holds this view. He prefers to explain these phenomena as extended phenotypes, which classifies them as external extensions of the biological gene instead of as separate entities with their own evolutionary accounts. However, evolutionary biologist Richard Dawkins warns against trying to explain everything through biological

evolution, and that the use of cultural evolution is essential in explaining certain phenomena, especially with regard to *group selection*, such as tribalism, nationalism, religion, etc. (Dawkins, debate with Weinstein 2018). 'Group selection' is an evolutionary concept that explains how groups (i.e. lineages or traditions) that share a common gene or meme are selected for over other groups, even if in cases where that gene or meme might be selectively deleterious to the individual (Dennett 2007, 106).

While both sets of theories rely on the same fundamental principles of variation (mutation), selection, and replication, they each are positioned to explain different types of phenomena. Some phenomena, Dawkins argues, are better explained from the point of view of a self-replicating meme, rather than a biological gene that is using the environment to ultimately better its own self-replication. Differences can be seen in the range of phenomena they explain, the speed at which they operate, and the apparatus that they require in order to replicate. One cycle of biological evolution requires a biological reproduction, or one generation, along with a mutation in the genome structure (this structure is typically quite robust and mutation events are limited (Dennett 2014, 22)). However, one generation of cultural evolution only requires the transmission of an idea, and a mutation in the idea (which is highly likely, considering ideas tend not to be as robust as genomes in the fidelity of their transmission). These differences create trade-offs that affect the replication in different ways.

A suitable example of this is a celibate catholic priest who does not reproduce any genes, but dedicates his life to the reproduction of catholic memes. This is, arguably, more beneficial to the evolution of the organisation because of the higher rate of transfer of memes versus genes. Considering that one male could perhaps pass on their 'catholic proclivity' gene to ten offspring, whereas he could transfer the catholic meme to hundreds of hosts over the course of his life. However, there is a trade-off between the strength of his hosts. Perhaps he only retains 10% of the hosts he converts with his memes, but would have retained 90% of the offspring hosts he had raised in a catholic household, he would have been better off having lots of children! But then another trade-off is introduced, once we consider how the meme can leverage biological evolution: considering that catholic religion discourages the use of birth-control, increasing the likelihood that meme-hosts will reproduce lots of children gene-hosts. Imagine that the 10% of this priest's meme-hosts

have children with a 90% gene host retention rate, so again, this priest is better to create meme-hosts instead of gene-hosts. It seems paradoxical if this scenario is examined only through biological theories, because the priest's actions are negative with regard to the catholic meme (i.e. the biological tendency toward belief) being transferred though genetic reproduction, however still positive for the overall reproduction of the meme. This example highlights that while cultural evolution may ultimately reducible to biological evolution, it should be examined at the memetic level for the sake of comprehensibility.

This subsection has established that evolutionary theories can provide empirical accounts of religions. Richard Dawkins promotes religions as being memetic in their evolution, whereas Bret Weinstein insists that they should be given a biological account. This distinction will be addressed in the main text of this thesis, where I will elaborate on which approach is more suitable depending on the level of complexity of the subject. Nevertheless, in either instance, a satisfactory account can be created.
Harvey Whitehouse and colleagues have also shown that an evolutionary account of a religion provides a framework to build a model of the phenomenon, which can be used to predict later developments. This feature will enable the identification of the origins and predicted developments of AI environments.

1.4 Artificial Intelligence, Techno-Social environments, and Take-offs.

Generally, the term ‘artificial intelligence’ is used to describe either of two methods of computation. In the first instance, the term is used to describe a computation that completes a specific task that would generally require intelligence to complete (i.e. playing chess or filing an appeal against a parking violation), also referred to as ‘narrow’ AI. Other terms for these computations include: ‘learners’ (as in, machine-learners), or ‘soft AI’ (Domingos 2017, 23).

In other cases, the term is used to describe ‘general’ AI: genuine intelligent cognition that is the result of computational processes (some definitions stipulate the inclusion of silicon chips or the exclusion of any biological processes, but such a definition would be superfluous for the needs of this review). General AI is also commonly referred to as: ‘real AI’, ‘hard AI’, or artificial general intelligence (AGI) (Tegmark 2017, 40).

1.4.1 General AI.

AI researchers have a very broad range of predictions of when they think we will create General AI, yet very few think it will never happen (Müller and Bostrom 2016).

The term ‘General AI’ describes a genuinely intelligent system that is capable of independent general cognition, essentially an ‘ability to accomplish any cognitive task at least as well as humans (Tegmark 2017, 39). A system with this capability would act with autonomy and have the capacity for robust problem-solving, extrapolation from existing knowledge, generalisation, and recursive self-improvement.

The term ‘General AI’ is, paradoxically, both specific and ambiguous. The term excludes algorithms that are explainable (i.e. symbolic reasoning) and only perform within a specific domain (if it only performs within a single domain, then it is not ‘general’), but is also ambiguous about what cognition actually constitutes and what classifies as ‘general intelligence’.

Intelligent problem-solving within a single domain is not sufficient for general AI. Once an algorithm is comprehensible, then it has been demystified and relegated to ‘narrow’ AI where it gets used for specific purposes that it performs well in. John McCarthy, a

founding member of the AI research field, says "As soon as it works, no one calls it AI anymore" (Bostrom, p.14).

There is an assumption in the AI field that once AGI is achieved that it will inevitably become super-intelligent, due to an accelerating loop of recursive improvements (i.e. an AGI that is as intelligent as a human could improve its own code, which would then continue to optimise until the only force that would decelerate it would be the laws of nature, such as the speed of an electric signal through a wire etc) (Tegmark 2017, 40; Bostrom 2016, 24, 90). Philosopher David Chalmers provides an insightful account of different acceleration arguments in his paper *The Singularity: A Philosophical Analysis* (Chalmers 2016, 3). This process would exponentially accelerate AGI from its base level of 'just like a human' intelligence to an inconceivable level of intelligence, often referred to as 'super-intelligence' (Bostrom 2016). The predicted timelines on this vary (Bostrom 2016, 24), but it is generally understood as being an inevitability. For this reason, and for the purposes of this thesis, AGI will be treated as synonymous with superintelligence.

In the context of this research, General AI, or AGI (artificial general intelligence) would make an ideal comparison to popular conceptions of 'god'. An AGI would be super-intelligent (perhaps even omniscient), and omni-present (through the 'Internet of Things'). It would most definitely 'act in mysterious ways', because–as has been speculated by AI researchers–its motivations would be completely incomprehensible to us (S. J. Russell 2019, 132; Bostrom 2016, 253; Tegmark 2017, 135). However, despite their likeness, ironically, a comparison between the two 'gods' would be nothing more than speculation, because of how unpredictable and uncontrollable the outcome could be. The unanimous decision is that it would dominate the human race with little regard for its welfare, although this is somewhat contested by Max Tegmark who proposes potential positive outcomes which feature AI as a 'Benevolent dictator', 'Gatekeeper', or 'Protector god' (2017, 162) (I use the term 'somewhat', however, because these possibilities are also accompanied by just as many negative potential outcomes).

In summary, the concept of super-intelligent AGI would provide a textbook comparative to a God, but it would lack substance in its applicability in the short-term, and lack accuracy in the long-term. It will not be addressed further in this research other than in the discussion around 'take-off scenarios' to be addressed later.

1.4.2 Narrow AI

The term ‘Narrow AI’ describes a category of algorithms that complete specific computational tasks within fixed domains. At the various times of each their individual conceptions, these algorithms were contenders for general intelligence, and caused much excitement in the field of artificial intelligence with their potential. Pedro Domingos explores this in his book, *The Master Algorithm* (2017). There are 5 ‘tribes’, as he describes them, each of which is a school of thought dedicated to certain approaches to algorithmic intelligence. Each ‘tribe’ believes that their machine learning method holds the potential for cognition, but for the most part these algorithms excel in certain domains and lack in others, none showing consistent promise of being ‘the master algorithm’ as researchers evaluate their ability to generalise in different environments. This limited domain of applicability is what is meant by the term ‘narrow’. The algorithm is used within a limited domain to complete what could be an extremely complicated task, but they are not generally intelligent and cannot complete tasks outside of the domain which they were designed for. Some AI researchers say that these are not ‘real’ AI and should only be considered as machine learning algorithms or techniques for statistical analysis, but we will not concern ourselves with the internal battles of the field.

The important aspect of narrow AI is that it is a field that is constantly growing and being researched. This continued growth is facilitated by the continuous search for ‘general AI’, which results in algorithms being evaluated and applied to complex problems in a variety of scenarios. Once these algorithms are found to work in certain contexts (and found not to work in others), they are adopted by various ‘narrow’ disciplines and replicated throughout their specific domain.

While narrow AI may not hold the grand title of being generally intelligent, these algorithms are actually much more widely adopted because they are comprehensible to those aiming to use them in practical application, and proven to work in certain areas, substantiated by comprehensive academic research. While this may be disheartening to the researcher looking for ‘the master algorithm’, it is very beneficial *to the algorithm* to move from general AI into narrow AI, because *an algorithm that is useful gets replicated*.

The proposal that prompted this thesis is inspired by the super-intelligence that would result from general AI. This proposal claimed that if we ever achieve general AI, then it would inevitably become super-intelligent (as previously discussed), which should therefore be worshipped because anything more intelligent than a human should be considered a god. However, religions do not literally *have* super-intelligent Gods, one can only guarantee that they *have the idea* of a super-intelligent God. Therefore, having a system that achieves super-intelligence is not necessarily a prerequisite for making a religion where artificial intelligence is the 'God', *the idea* of it will suffice.

This research will purposefully not use examples of general AI in the comparison between AI and religion, because its existence is too reliant on unpredictable outcomes, and as explored; not a necessary prerequisite for a religion. To include general AI would make the comparison much easier, but much less robust, as it would only have implications for extreme scenarios, where insight would be mostly redundant given the drastic societal and behavioural changes that would occur instantaneously (discussed later in 'take-off scenarios').

1.4.3 AI Environments as distributed systems:

This research will use the term 'artificial intelligence' to describe collections of narrow AI algorithms that play an active role in human decision-making – which will be referred to in this research as 'AI environments'. The algorithms in AI environments may change, they could be upgraded or deleted, more could be added, or they could be reduced to one 'master algorithm'. What qualifies them as being 'artificial intelligence' in this research is that these systems quantify human behaviour in some way (through some type of representation such as counting steps or minutes spent reading an article), manipulate it (pass it through an algorithm which has variable weights that can be trained), and provide some output that is not hard-coded (the response must not be the result of a defined set of explicit rules). These systems may be embedded within complex environments, such as an actuarial algorithm that calculates life-insurance premiums based on observations on wearable technology, health data, demographic profiling from social media accounts, and court history, or these systems can be high-level with only one level of quantification, such as a recommender system in an online content provider. This is not to say that each of

these systems could be separate religions, but that they are all contained under the general concept of a unified 'artificial intelligence' used in this research.

There is an appropriate comparison here with a common analogy used to explain how God can be father, son and holy spirit (the holy trinity): it is like the light in a room with a window, a ceiling light, and a lamp. The light (God) does not originate from just one source, and this light cannot be divided into its constituent parts, however if you were to take away all three then it would disappear. It is not limited to one light source, but it is dependent on their existence to a degree, in the sense that it is constructed by them. While this is not an exact comparison to what I am describing under 'artificial intelligence' because my definition encapsulates a full comparison with every aspect of a religion (not just a comparison between AI and God), it demonstrates how a sub-collection of algorithms could play the role of a god-equivalent in an AI environment, without having to be *all* of the religion.

The concept of an 'AI environment' that I will be introducing in this research resembles something close to Shoshana Zuboff's 'Big Other' (2019, 376), and Brett Frischmann and Evan Selinger's 'Techno-Social Dilemma' (2018, 9). Both of these concepts are (or could be) facilitated by narrow AI algorithms, but cannot be defined by any specific algorithm. Zuboff describes her 'Big Other' (inspired by 'Big Brother' from Orwell's *1984*) as 'the sensate, computational, connected puppet that renders, monitors, computes and modifies human behaviour', it 'reduces human experience to measurable observable behaviour while remaining steadfastly indifferent to the meaning of that experience' (376). Zuboff's description implies the system as being something that is intentionally used ('puppet'), which poses an immediate issue because it makes it hard to evaluate this system objectively when its use has a defined objective.

AI environments typically have some sort of public facing front-end where users communicate with them. For instance, a percentage recommendation score on Netflix, or an input form and output quote on an insurance site, or a spoken recommendation from a smart home-assistant. Importantly, there is a popular understanding that these front-ends are the public faces of 'the algorithms'. In a sense, it does not matter if these computations are strictly powered by AI (narrow or general), but rather there is a marked perception (and frequent claims) that they are. With this in mind, these AI environments will largely be

discussed in the context of social media platforms and recommendation engines. The reason for this being that these are the most frequent area where the general public enter AI environments. These scenarios are used as *examples* of interactions between users and AI environments, rather like observing religious believers attending a mass or confession. They should not be seen as the *extent* of AI environments, but rather relatable examples of behaviour when interacting with them, which I believe can be applied generally to other scenarios. In much the same way that a theologian may document and illustrate the beliefs and culture of a religion through analysing their behaviour in various ritual and everyday acts.

1.4.4 Separating the system from the owner.

To separate the hierarchy of owner and system identified in the previous subsection, it is necessary to interpret Zuboff's description using Daniel C. Dennett's 'intentional stance'. Dennett's intentional stance is a 'strategy of interpreting the behaviour of an entity… by treating it *as if* it were a rational agent' (2014, 78). This means interpreting behaviour through beliefs and desires (intentions) as the easiest way to understand the behaviour of the system (as opposed to trying to understand exactly how it is designed at an algorithmic level (what Dennett calls the 'design stance' (80), i.e. how it is designed), or even the more complex physical level where you would have to observe electricity passing through transistors (what Dennett calls the 'physically stance' (79), i.e. how it physically operates). To understand Zuboff's description using the intentional stance is to understand that it is not that the system *wants* to 'reduce human experience', but that this is the best way to describe how it accomplishes what it is designed to do (which in the case of Zuboff's context, is maximising profit through behavioural modification). Using the intentional stance, Big Other can be reduced to being understood as a process of quantifying human behaviour and the *desire* within the system (the puppet) to provide an appropriate response based on an implicitly trained value-system (the user's needs). This value-system could be trained to find the optimal combination of profit and human satisfaction in the case of a recommender system that tries to find the cheapest movie a viewer would enjoy based on their viewing history (evaluating a matrix of provider cost and consumer satisfaction). Alternatively, this value-system could be trained on an authoritarian value-system that uses

facial recognition to identify jaywalkers and automatically deduct a fine from their bank accounts (evaluating a matrix of federal law and individual liberty). The need to adopt the intentional stance in this context is because these value-systems may not have been explicitly taught to the algorithm (as would have to be the case if it were a 'puppet'), but could be implicit in the environment in which it is trained. This can be understood like a guard dog who is a 'puppet' of the owner. It has been trained to attack intruders, and you could say that the dog *wants* to attack intruders. But it is not that the actual underlying functions of the dog want that (the actual algorithms; jaw functions, leg muscles). The functions that facilitate 'attack' have been hijacked for the desired purpose of the owner. In much the same way, the functionality of Big Other has been hijacked for the desired purpose of the owner, similar concepts to which are explored in Safiyah Noble's *Algorithms of Oppression* (2018) which identifies the implicit ideologies of engineers being reinforced in the algorithms being created. Now that a value-system has been introduced, this aspect of the system must be considered as crucial to the overall survival of the system. If the system doesn't *want* to do the things the owner wants it to do, then it won't get used and won't get replicated. This can reduce being a 'puppet' to just a selective constraint in the environment that it needs to survive in, rather than being its *raison d'etre*.

Applying the intentional stance helps to understand algorithms (or systems of algorithms) in evolutionary terms, in the same way that Dennett uses the intentional stance to understand the 'motivations' of an organism or idea with regard to their (and subsequently, our) evolution (2014, 171).

The separation of owner and system means that this thesis can focus on the functions of the system in evolutionary terms, rather than evaluating the evolutionary strength of the circumstance it is used in. For example, take an AI system that profiles criminals on whether they will re-offend, the output of which is used to judge whether the prisoner gets put on parole. This system *wants* to perform well, because then it will be reproduced (i.e., it will *try* to give correct answers, which is the intentional equivalent of saying it will be designed with a well-trained algorithm with a high accuracy score). Systems that do not *want* to perform well (i.e., are *not* designed with a well-trained algorithm) will not *try* to give correct answers. In this way, we can give an evolutionary account of the system,

without the need to address whether predicting criminal profiles is an evolutionarily beneficial thing to do.

As a comparison, recall the trained guard dog. If the guard dog is good at guarding because he *wants* to make his master happy, then we can give an evolutionary account for this behaviour, regardless of what constitutes the master's happiness (in this case it's an effectively guarded residence).

This will be relevant to this research because it will give an evolutionary account of what features make AI environments 'strong', rather than evaluating the 'strength' of what they are doing in the specific context they are currently used in (because that context can change).

1.4.5 AI environments embodying a value-system

The previous subsection introduces a crucial aspect of these environments, insofar as they temporarily embody an underlying, implicit value-system, like the trained dog temporarily embodying a value-system of being aggressive. This can be seen as an interpretation of an absolute order of the universe, at a fixed moment in time. That is to say, the system 'knows' what is 'right' and 'wrong', and will prescribe action in accordance with that set of values, because it 'wants' to perform well. This will be addressed in more detail and in the relevant contexts throughout the thesis.

1.4.6 AI environments as an unintended by-product.

Frischmann and Selinger's term 'Techno-Social Dilemma' describes something less subordinate than 'Big Other'. They use this term to describe the issue humans face with regard to 'techno-social engineering'. Techno-social engineering encapsulates the idea that humans are 'being conditioned to want to obey' (Frischmann and Selinger 2018, 6) by the machines they use. Technology is typically considered to enhance human experience and work *for* humans. Techno-social engineering suggests that this may not be the case, and that 'our preferences are increasingly manufactured rather than freely adopted, thanks to techno-social engineering calling the shots' (6). This portion of the theory resembles the underlying message in Zuboff's warning about 'Big Other', however Frischmann and

Selinger emphasise strongly that these developments are driven by rational choices that make sense within the context that they operate, but overall have a negative effect in the wider environment. They liken this to the 'Tragedy of the Commons'. The Tragedy of the Commons is an environmental allegory from ecologist Garrett Hardin, in which a community of sheep herders have communal access to a pasture of land. Each herder uses the common to feed their sheep. 'And so, each individual proceeds under the assumption that it's rational to increase the size of her own herd to capture the benefits of a pasture that everyone shares while only bearing a fraction of the costs that accrue as the common resource gets exhausted' (Frischmann and Selinger 2018, 9). In this scenario each member of the dilemma makes small, rational decisions, but the aggregation of which results in disaster. The solution, the authors stress, is that the members in the dilemma need to 'better understand their relationships to each other and their shared resources and develop governance strategies for cooperatively bringing about sustainable well-being' (9). This is compared with the current dilemma they claim we are facing with techno-social engineering, in which individual decisions regarding technological developments are rational, but the overall impact could result in us 'rely[ing] on the techno-social engineers' tools to train ourselves, and in doing so, let ourselves be trained' (10). This, they say, is '*humanity's techno-social dilemma*'.

These concepts have a relevance to this research because they are not reliant on one specific technological development, in the same way that religions are not reliant on one specific idea of god. In fact, these concepts could potentially exist without the need for any artificially intelligent algorithm, however these algorithms greatly facilitate the accelerated adoption of these systems through increased automation and an increased ability to accommodate an individual's behaviour. In comparison to religions, it may not be necessary to have a god figure (as explored earlier while dissecting Dennett's definition), but perhaps it greatly accelerates the adoption of a religion once one is introduced.

The most interesting aspect–and most relevant to this research–is that these systems originated as distributed, disconnected, unorganised practicalities whose overall influence is far greater than any of the individual algorithms or practices. Whether these systems are used as 'puppets' like Big Other, or unfortunate by-products of 'the rational behaviour of

producers and users who develop, deploy, adopt and use innovative technologies' (6) like the techno-social dilemma, is to be explored in this research.

A comparison can be drawn between the systems of religion and the systems of 'artificial intelligence' if this conceptual framework is adopted. This research will explore the original purposes and by-products of these distributed systems, and then use theories of biological and memetic evolution to examine the centralised systems that form when these features are combined.

1.4.7 Take-off scenarios.

The terms 'fast take-off' and 'slow take-off' are introduced in Max Tegmark's book *Life 3.0*. These terms are used to describe the two categories that a potential 'intelligence explosion' may have.

An 'intelligence explosion' (Tegmark 2017, 39) describes the way in which an AGI (artificial general intelligence) system would take control of society. The assumption in this scenario is that if an AGI became more intelligent than us, that it would dominate us, in the same way that we dominate other lifeforms due to our superior intelligence.

A fast take-off is a scenario where one entity obtains a monopoly over a technological breakthrough that creates a super-intelligent system. In this scenario, the owner of the system (which could be the system itself, if it has a 'personal' motivations) gains very significant monopoly over the new technology and its capabilities. The likelihood that this type of technological breakthrough would be discovered by two disconnected entities at exactly the same time is highly unlikely, and therefore a monopoly over the new technology is significantly likely. The speed at which a super-intelligent system could make recursive improvements on itself would only stand to significantly increase the gap between this entity and its opposition at an exponential rate that there would be little chance of even a fast-follower.

Given the unprecedently levels of potential intelligence available, Tegmark does not need to utilise a lot of imagination in describing numerous methods that a super-intelligent system would use to take control over our current economic and political systems of power. He describes scenarios that are achievable by humans, provided there were enough

of them all aligned on a single outcome on a large-scale (generating funds on Mturk, creating shell corporations, selling video-games, etc) (9).

A fast take-off is highly undesirable, because of its potential to disrupt systems of power in such a short time-frame. However, this scenario is highly unpredictable too. While Tegmark uses common sense to create plausible scenarios, the entire endeavour cannot be considered more than speculation at this stage.

A slow take-off describes, instead, the gradual integration of AGI with society, which has consistent but incremental effects on power-structures. It does not describe a scenario where one entity has a monopoly over super-intelligence, instead it identifies ways that artificially intelligent systems could affect power hierarchies in a more distributed way that would not be as disruptive to the *Nash equilibrium* (a concept from game theory, in which each party is incentivised to cooperate to everyone's gain. Tegmark uses an example of everyone benefiting by sacrificing some of their power over to a government, but also in turn retain collective power over that same government (151)). We can safely speculate that a Nash equilibrium would not be maintained in a 'fast take-off' scenario, as all power is allocated to a single party almost immediately. In a slow take-off scenario, there is a significant focus on individual and distributed changes, enabled by AI, that have incremental impacts on power hierarchies. Some of the possible changes that Tegmark highlights occur in (but are not limited to) transportation, communication, biology and intelligence (in the form of cyborgs), surveillance, and government relations. While developments in these fields will not disrupt the power equilibrium over-night, we can imagine how incremental changes could still have significant implications for power hierarchies. To examine how AI could integrate with–and have an impact on–existing power hierarchies over the course of a slow take-off affords us the chance to ensure AI is used as the tool it is meant to be.

It is important to note that Tegmark discusses these take-off scenarios relative to AGI. This need for General AI clearly applies in the fast take-off scenario, as Narrow AI would not be capable of the intuition and self-improvement capabilities required. However, the slow take-off scenario is also plausible for Narrow AI systems because the scenario relies on humans using intelligent systems (which is what Narrow AI is intentionally built to be),

not on intelligent systems using humans (which is what AGI is anticipated to do). The presence of General intelligence in a slow take-off scenario would indeed make it happen faster, but I do not think it is a necessary condition in the development of a society dependent on artificially intelligent systems.

1.4.8 Summary of the technological context of this research:

This research will focus on techno-social environments that are built on collections of narrow versions of AI within a slow take-off scenario, for two reasons:

1) If collections of narrow AI can become a religion, then we can guarantee that general AI can be (it will have everything that soft AI has, plus genuine cognition).
2) Focusing on narrow AI gives us the best opportunity to anticipate and integrate AI into our lives in a healthy way throughout a slow take-off. Additionally, these algorithms already exist and their capabilities do not require any speculation.

By using Dennett's intentional stance, AI environments can be described in evolutionary terms, without the need to evaluate the evolutionary strength of the context they are used in. This is necessary to understand how or why it could originate, how successful it will be adopted, and how it will integrate with existing systems (both technical and social).

Importantly, however, we must acknowledge that these systems are used in various contexts, and that the system will embody the value-system of the context it is placed in. This will have implications for examining the way it will prescribe and shape behaviour within the broader system it plays a role in.

1.5 Philosophy of Mind

This research will explore the consequences that outsourcing decision-making to external systems (e.g. AI environments) would have for an individual. These consequences relate to philosophical theories of the mind and will examine how an individual could determine their own values and maintain moral responsibility in scenarios where decisions are potentially being made *for* them (as opposed to *by* them). Immanuel Kant believed that self-determination of the will was absolutely essential to moral responsibility and freedom, and that one must absolutely avoid adopting moral values that were formulated on insecure foundations, such as religious ones are. As Will Durant summarises in his chapter 'Kant and German Idealism': 'the moral basis of religion must be absolute, not derived from questionable sense-experience or precarious inference; not corrupted by the admixture of fallible reason; it must be derived from the inner self by direct perception and intuition' (Durant 1927, 300). Adopting a Kantian view of moral responsibility, we can know we are free by 'feeling it directly in the crisis of moral choice' (Durant 1927, 302). While this 'crisis' might be reduced in a scenario where one's moral choices are made on their behalf, we can conclude that the choice being made must be 'derived from the inner self' (i.e., self-determined) as a necessity for individual freedom.
This review will aim to establish if humans can retain mental individuality (as synonymous with individual free-will) in instances where decisions are being made for them.

These theories will further reference empirical studies in the fields of cognitive science and behaviour psychology.

There are a number of philosophical assumptions made in the paragraph above that must be explicitly stated and evaluated in order to establish philosophically sound arguments within the overall thesis.

Firstly, there is the assumption that the mind is a result of–but also not limited to–processes in the brain. This will be addressed through an exploration of 'physicalism', 'functionalism', and 'multiple realisability'. Additionally, the idea of 'individuality' referenced above assumes that the mind controls decisions. This assumption will be addressed by exploring neurological research into the control the brain has over decisions, which will imply that the individual could choose to do otherwise. The assumption that if someone can do otherwise means they can determine their own values and should be

morally responsible for them (i.e., how could an individual who is in control of their decisions not be free?). As we will see, Harry G. Frankfurt disproves the common assumption that a free choice requires the ability to do otherwise, which will be addressed through an investigation into theories in the philosophy of free-will, referring to determinism and compatibilism. This will provoke the need to establish a definition of 'freedom', which will be constructed from ideas central to Kant's view on individual freedom and the necessity of self-determination. Once these assumptions are established, then we can conclude that freedom, moral responsibility, and self-determined values are dependent on physical processes in the brain which are multiply realisable, and could therefore be replicated in part by a machine. This claim will then be evaluated with reference to Clark and Chalmers' 'The Extended Mind' theory, which asserts that a mind can be extended (in part) to a machine.

1.5.1 The mind as a result of, but not limited to, processes in the brain.

In order to maintain the naturalistic epistemology of this research, it is necessary that mental states be recognised as the result of some physical process, such as processes in the physical brain. This position is known as 'physicalism'.

Functionalism is a concept in the philosophy of mind that states that the mind (i.e. mental states) should not be defined by 'its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part' (Levin 2018). In this respect, the idea of a mind-body dualism is not necessarily excluded as a valid source of mental processes, provided the non-physical substance could play the correct functional role. However, that need not concern this research, because the same concept can be used in a physicalist approach, in which the concept of functionalism also maintains that mental states can be understood as being the sole result of any physical processes and need not rely on any form of dualism (provided the physical processes can play the same functional role). This facilitates a naturalistic interpretation in which mental states can be understood as being the results of physical mechanisms in the brain, but not equivalent to the brain itself, and also not limited to *only* being results of processes in the brain. This will be a requirement if we are to say that a decision made within a computer can ever be considered as being made *by* an individual, instead of *for* them. This plasticity in embodiment is

referred to as ‘multiple realisability’, a thesis that ‘contends that a single mental kind (property, state, event) can be realized by many distinct physical kinds’ (Bickle 2020). That is to say, that a mental state can be ‘realised’ by means of brain processes, cogs and wheels, or silicon chips, provided they all maintain the exact same functional properties.

It should be noted that the term ‘functionalism’ can be used to define other concepts in separate fields (e.g the functionalist approach in mythology, which will be encountered later) and should not be interpreted as in any way linked to its use in this area.

1.5.2 The mind retains executive control and the ability to do otherwise.

Functionalism allows us to state that mental states are the results of physical processes in the brain (or any other physical process, for that matter), explicable through their functional properties, and that this is potentially extendable to other physical processes, provided they maintain the same functional role. In this way, we can extrapolate that the decisions could legitimately be made outside of the *brain*, but still be considered as being made in the *mind*. This concept does, however, lack a sense of self or individualism.

Experiments in neuroscience indicate that decisions may be made by the brain prior to a conscious awareness of them (Schultze-Kraft et al. 2016). This is referred to as ‘readiness potential’ (RP), where the brain is observed as having a ‘buildup of electrical potential … before a movement’ even before ‘the conscious awareness of the decision [to move]’ (Fifel 2018, 784). However recent studies have ‘demonstrate[d] that premovement RP is not sufficient for the enactment of a motor action’ (786). Therefore, we can maintain that at the very least, humans always retain the ability to stop an action to some degree. Schultze-Kraft et al. describe a ‘point of no return’ (2016, 1080) 200 m/s before the action where the action can no longer be vetoed by the individual, however Fifel references another study that found that subjects maintained the ability to ‘alter and abort the movement as it unfold[ed]’ (2018, 786). With this information, we can determine that decisions made by the apparatus that also make mental states are within ultimate executive control of the individual, even in cases where they are not necessarily consciously initiated by the individual.

This is also not to say that all decisions are made unconsciously by the brain, and that the only ability to determine values that the individual retains is the ability to veto or allow decisions made elsewhere (which would not constitute a self-determined decision). Of course, conscious decisions are viable sources of decisions, but this review aimed to investigate whether *all* decisions can be attributed to the individual, even in cases where they originate from an unconscious source.

This ability to create conscious decisions or approve/cancel unconscious decisions constitutes the ability to *do otherwise.* The ability to 'do otherwise' is a contentious feature in the philosophical debate regarding the compatibilism of determinism and free-will. This is a concept that I do not wish to address in much depth in this review other than to acknowledge that the ability to 'do otherwise' constitutes free-will on both sides of the debate, however incompatibilists say that if we live in a deterministic universe, then we do not retain the ability to 'do otherwise', whereas the compatibilists say that in a deterministic universe we do (Beebee 2013, 9). For the purposes of this research, we will assume we do live in a deterministic universe (to rule out cases where randomised events can occur), and that we are adopting the compatibilist view (to avoid arguments about a deterministic universe stalling arguments about free-will that will arise later in the thesis).

1.5.3 Self-determination and moral responsibility as sufficient for individual free-will.

In his paper 'Alternate Possibilities and Moral Responsibility' Harry G. Frankfurt shows that the 'principle of alternate possibilities … [which] states that a person is morally responsible for what he has done only if he could have done otherwise' (1969, 829) is false. Frankfurt uses an example of a nefarious neurosurgeon to illustrate his point. A distilled version of the thought experiment is that there is a man, Jones, that is contemplating killing another man, Smith. There is a third character, called Black, who is a neurosurgeon, and who wants Jones to kill Smith. Essentially, Black installs a mind-control device in Jones' brain, however it lays entirely dormant until Black activates it, with no residual side-effects. If Jones decides **not** to kill Smith, then Black will activate the device and Jones will reconsider and decide to kill Smith as a result. In this sense, Jones

cannot do otherwise than kill Smith, however it appears intuitively that he only retains moral responsibility for the killing in the case where he chose independently to do it, and not morally responsible in the case where he was coerced. This, Frankfurt claims, proves the Principle of Alternate Possibilities to be false because we intuitively accept Jones as acting freely in a situation where he could not do otherwise.

Throughout the paper Frankfurt alludes to the replacement of the Principle of Alternate Possibilities (PAP) with a similar principle that does not ignore the fact that a person could still 'bear full moral responsibility for performing [an] action… even though [that] person is subject to a coercive force that precludes his performing any action but one' (1969, 834) because that person could have still chosen to act in that way, as we saw in the case where Jones performed an action freely, despite being unable to do otherwise under a coercive force.

In her book *Free Will: An Introduction*, Helen Beebee formalises the replacement of PAP, which she calls the 'Principle of Unforced Action' (PUA) (2013, 141), which declares that it must not be the case that the person performed the action *only* because they could have not acted otherwise. Essentially, a person is not morally responsible for an action if they did not have the option to do otherwise, and also if they did not desire to do it. If they did desire to perform the action, then they are morally responsible, even if they had no other options available to them.

In the terms of this review, because the mind retains executive control over decisions, we can say that the mind (the individual) is morally responsible for its decisions because the term 'decision', by definition, implies a choice of at least two possible scenarios, so there will always be the option to have 'done otherwise', which coupled with the presence of an executive controller and the act of choosing establishes a determination of value (i.e. a desire of one option over another). Provided the individual retains executive control over the final evaluation of a decision, which essentially constitutes a declaration of value (i.e. desire/preference), then we can say that the individual retains moral responsibility for decisions made *by* or *for* them.

1.5.4 An individual that can determine their own values and be morally responsible for them is free.

Exercising the ability to self-determine values (i.e. to be able to think for oneself) is a necessary condition for the Kantian view of freedom (Durant 1927, 300–302). Kant strongly opposed the idea of an individual adopting a value that was not generated–or at very least evaluated–by the individual. As Bertrand Russell summarises: 'The essence of morality is to be derived from the concept of law; for, though everything in nature acts according to laws, only a rational being has the power of acting according to the idea of law; i.e. by Will' (B. Russell 2010, 644). This introduces Kant's 'categorical imperative' which states that one should 'Act only according to a maxim by which you can at the same time will that it shall become a general law'. This imperative is *a priori*, in the sense that it does not rely on any prior knowledge and is therefore achievable in any circumstance by any rational being. It is also categorical, in the sense that it 'is objectively necessary, without regard to any end' (B. Russell 2010, 644).

Kant issued this categorical imperative as the only way of achieving a morality that was not rooted in the foundation of something other; i.e. not an imperative to an other's end. While moral responsibility was always placed on the individual (i.e. the responsibility to follow the set of prescribed values), this placed a different type of moral responsibility on the individual, the responsibility to *decide* the values. This formulation came from Kant's belief (from his *Critique of Pure Reason*) that 'the objects of faith–a free and immortal soul, a benevolent creator–could never be proved by reason' (Durant 1927, 299), which necessitated that religions would need to be grounded on morals, seeing as it could not have a foundation in science or theology, and that the morals must be absolute.

Kant's categorical imperative provides a way for individuals to retain (or rather, attain) moral responsibility in the absence of religious prescriptions, rooted in the conviction that the truth of externally generated morals cannot be proved by reason, and therefore are likely created toward some other's end. Kant's approach to moral responsibility necessitates the self-determination of values, and through this a means of achieving

individual freedom, as the ability make moral choices.
This constitutes the definition of freedom that will be used within the scope of this thesis.

1.5.5 Is an outsourced decision an extension of the mind?

Humans are currently equipped with the tools necessary to uphold Kant's categorical imperative, which are: a mind (theoretically) capable of self-determining values, coupled with the ability to be held morally responsible. As this thesis will examine later, the mind may have the functional capacity to be free, but the human *mentality* may not have the operational capability just yet (referring to the religious mentality previously mentioned, which will be addressed in more detail in section 4). Take, for instance, the computer I am currently sitting at: it has a keyboard and mouse, and an internet connection. If I was to operate the keyboard in such a way it is theoretically feasible that I could get one million euros into my bank account, that is, I have the functional capability (tools) to do so (I have the computer, and a bank account, and I could get the money by either earning it by setting up an online business, hacking someone and stealing it, or receiving the money as ransom in some sort of blackmail scam etc). Unfortunately, I do not have the mental sophistication to do any of these, but that could be developed over time and eventually, perhaps, I would be operationally capable of the task. That is to say: I currently have the functional capability, but not the operational capability, yet.

In this sense, we have established that humans have the tools necessary to become free (i.e., a mind capable of self-determining values and being morally responsible for decisions) and they just need to work on the capability (to stop adopting values instead of generating them from within), however, if decisions are outsourced to an intelligent machine, then the necessary question to answer is: can we legitimately say that the decision is being made *by* the individual, or only ever *for* the individual? If the answer is 'by', then humans retain the capacity to be free in the Kantian sense, and (still) just need to work on the capability (for which machines may provide necessary assistance). However, if the result is that decisions are being made 'for' humans, then perhaps humans will lose the capacity–along with any potential capability–to be self-determining, morally responsible, free agents.

Philosophers Andy Clark and David Chalmers explore the concept of the 'extended mind' in their 1998 paper, 'The Extended Mind'. In this paper, the authors propose that cognitive processes are not restricted to the individual's brain and that the mind can exist as a 'coupled system' with the environment it is in (1998, 7). To illustrate their point, they use an example of a person playing Tetris and rotating a shape around 90 degrees using a button instead of performing the transformation in the foreground of their imagination. In this instance, the cognitive task of rotating the object is outsourced to the machine, however the authors would suggest that the phrasing used there was superfluous for explaining the action. They say that these external cognitive processes 'demand … *epistemic credit*… [because] were it done in the head, we would have no hesitation in recognizing as part of the cognitive process' (8), and therefore we should refer to the person's mind as having been extended.

For the purpose of this review, this example sufficiently illustrates the core idea that Chalmers and Clark propose. The extended mind will be explored in more depth within the context of AI environments in Section 4 of this thesis. For the time being, it is only important to note that if the extended mind hypothesis is accepted then we can legitimately say that decisions made by a machine in a coupled system can constitute an extension of the mind in that system, which would necessitate it being a self-determined and free decision, under the assumptions previously examined.

1.6 Conclusion.

This conclusion will summarise the four areas addressed in this section, which were:

- Religion
- Evolution
- Artificial Intelligence
- Philosophy of Mind

1.6.1 Religion.

The texts explored in this subsection on religion were used to create a working definition of religion for the purpose of this thesis, with specific regard to identifying early religious concepts and distinguishing a religion from a sense-making tool. Daniel Dennett's summary marked the conservative end of the hypothetical scale of definitions of religions, and Yuval Noah Harari's marked the most liberal. Dennett's definition could identify the major world religions as they appear at present day, but does so at the cost of being able to identify religions at their early stages. Harari's definition allows the categorisation more or less at the inception of any set of shared values, but does so at the cost of being able to identify the concepts that differentiate religions from sense-making tools. For the purpose of this research it was necessary to be able evaluate concepts as potential religions at their inception (perhaps currently without supernatural agents), and to not wrongfully classify sense-making tools (such as evolutionary biology or Wikipedia) as religions.

With these criteria in mind, the working definition for this thesis came to sit in between these two definitions as: *any concept–or collection of concepts–that prescribes a value-system that shapes human behaviour, that bridges a gap in explicability and irreducible complexity through requiring belief in an entity whose abilities are incomprehensible to the human mind and encourages a reluctance to further critical inquiry justified by a representation of an absolute order of the universe*.

Texts from Alan Watts, Carl Jung and Yuval Noah Harari also established a sharp distinction between spirituality and religion, which allowed the former to be removed from any further involvement within the research.

1.6.2 Evolution.

This subsection documented examples of researchers using theories of evolution to explain religious phenomena. The texts that were used as examples were philosopher Daniel Dennett's *Breaking the Spell*, evolutionary biologist Richard Dawkins' *The God Delusion*, and anthropologist Pascal Boyer's *Religion Explained*. This demonstrated that using evolutionary theories to explain phenomena is not reserved to the field of evolutionary biology, but rather a suitable naturalistic approach that can be used to explain this type of concept in an empirical way.

The second half of the subsection examined which strands of evolutionary theory (biological or cultural/memetic) should be used at the various points of the thesis. Bret Weinstein asserted the view that religions should be considered as extensions of the human gene ('extended phenotypes') and should therefore should only be evaluated through Darwinian theories (biological evolution) as relevant only in the context of the evolutionary fitness of the human 'host'. Richard Dawkins, however, defended that religions should be examined as entities in and of themselves, as ideas (memes). Dawkins maintained that this approach afforded benefits in explicability and simplicity, rather than having to channel everything through the human gene in biological terms. Further exploration established a common ground between these two theories, where Weinstein's would be used to explain the origin and early stages of religions, before they had become entities and were best understood as extended phenotypes, and Dawkins' approach would be adopted later when religions began to depend less on the biological fitness and adaptation of the human gene, and came to depend more so on human culture and tradition.

This subsection established theories of evolution as an appropriate positivist approach to explaining religious phenomena, and it also established the different strands of evolutionary theories to be adopted depending on the religious phenomena to be explained.

1.6.3 Artificial Intelligence.

The subsection addressed two aspects of artificial intelligence (AI) with respect to this thesis: what it is (and could potentially be) and how it will integrate with society.

The first portion of the review outlines the different types of artificial intelligence, the first being ‘general AI’ informed by Nick Bostrom’s *Superintelligence* and Max Tegmark’s *Life 3.0*, which is contrasted with ‘narrow AI’ informed by texts from Pedro Domingos and John McCarthy. An emphasis is placed on the fact that intelligent systems need not be limited to just one ‘master algorithm’, but rather are likely to be a collection of smaller algorithms performing specific tasks. This non-necessity of a master algorithm is further developed by concepts from Zuboff’s ‘Big Other’ and Frischmann and Selinger’s ‘Techno-Social Dilemma’, where ‘master algorithms’ are not required, and later used to define the ‘AI’ portion in concept of an ‘AI environment’.

The potential options for the societal rise and integration of AI environments are informed by the same texts from Zuboff, and Frischmann and Selinger. Zuboff’s ‘Big Other’ represents an emergence of these systems as ‘puppets’ of a conscious guiding hand, whereas Frischmann and Selinger’s ‘Techno-Social Dilemma’ is conceptualised as an unfortunate by-product, much like a digital ‘Tragedy of the Commons’. It is established that AI environments should be examined as emerging as unintended by-products, because it is critical that these should be identified early in their inception, and would be harder to detect if unintended by-products rather than as conscious decisions made by controlling actors. This approach is compared with similar explanations of the emergence of religions as unintended by-products, as is suggested by Daniel Dennett and Richard Dawkins, which provides a foundational framework in which AI environments can be evaluated.

The final aspect of the emergence of AI environments to be addressed is with regard to ‘fast’ vs. ‘slow’ take-off scenarios from Tegmark’s *Life 3.0*, from which the conclusion is drawn that the scenario that facilitates the most robust and actionable research is to expect a slow take-off, with narrow AI, in which an AI environment is an unintended by-product.

1.6.4 Philosophy of Mind.

The theories reviewed in this subsection determine whether it is feasible for humans to retain freedom, individuality, and moral responsibility, in cases where decisions are being made on their behalf by an intelligent system. Immanuel Kant’s categorical imperative and

views on self-determination were explored in order to provide valid reasons for why we should desire to retain free-will, individuality, and moral responsibility.

In addressing the implications of outsourcing a decision to a machine, adopting a functionalist approach to the philosophy of mind allows us to expand our interpretation of viable mind-facilitating-substances beyond biology, provided the mind-substance plays a sufficient functional role. For the purposes required, functionalism holds too broad a scope, in so far as it states that a mind can be realised in any medium. The theory is later contextualised within the scope of this thesis by Clark and Chalmer's theory of the 'Extended Mind', which promotes the view that the mind can be extended (but not in its entirety) to an external system. This strand of the discussion concludes that a mind can be extended to another non-biological substance, however Clark and Chalmers create a necessary bound that it cannot be *fully* extended, which is accepted, considering that within the current scope we only want to evaluate an extended mind, not a replicated or detached mind.

To address Kant's requirements for the categorical imperative, reviewing neurological research on readiness-potential provided empirical evidence that individuals do retain executive control over unconscious decisions, which implies that humans retain the ability to 'do otherwise' and to determine value. In questioning whether this necessarily constituted the decision as belonging to the individual (i.e. they were responsible for the decision) Frankfurt's example of the nefarious neurosurgeon refuted the 'ability to do otherwise' as sufficient for moral responsibility, however Helen Beebee's additional 'Principle of Unforced Action' established that moral responsibility is maintained provided that the individual is capable of *wanting* to do otherwise. This–combined with the previously mentioned 'Extended Mind' theory–identified the requirement for some type of executive controller within the 'couple system' of individual and machine, as a fully outsourced decision would not leave the individual capable of wanting to do otherwise (because they would be unaware of the decision being made or any other options available).

The two streams of thought were then combined to summarise that an individual could potentially maintain free-will, individuality and moral responsibility when a machine

makes a decision on their behalf, provided that the necessary conditions outlined above are met.

1.7 Conclusion of Literature Review and Introduction to the Research

The thesis of this research is that AI environments will use existing evolutionary mechanisms that were also used by religions in order to evolve. These mechanisms typically have biological origins, however a portion of these mechanisms were further adapted or amplified in the cultural evolution of religions. Religions encouraged, curated, and leveraged a specific mindset that has not disappeared despite humanity's move toward a secular society. This research explores whether the religiously-primed mindset could attempt to fill a cognitive void with artificially intelligent (AI) environments in a post-religious society. To paraphrase David Foster Wallace: 'The person who jumps from a burning high-rise does not do so because they have overcome their fear of falling' (1997, 696), and likewise humanity has not effectively discarded the religious mentality recursively propagated across cultures for millennia. In the West, (where these AI environments are prevalent), people have only suppressed their religious mentality, largely deeming it of no more use. These people have only just jumped out of the metaphorical window of religion. This research explores the remarkably similar features shared by AI environments and religions, and evaluates whether the former are likely fill the empty air that has been leapt into.

Section 2 begins with an exploration of the foundational origins of religion as by-products of early human thought, in the form of archetypes and myth. This is compared with the foundations of AI environments, which are behavioural datasets and profiling algorithms.

Section 3 will demonstrate how AI environments stand to capitalise on existing biological features that were also capitalised on by religion. This section will also explore how these biological features were further intensified by generations of recursive reinforcement by religious concepts. This research will outline a set of cognitive traits that were instrumental in the proliferation of religious ideas, which can be evaluated as potential mechanisms for the replicative strength of AI environments. These comparisons will provide context for a

later evaluation as to how AI could use similar mechanics to proliferate, ultimately exploring the probability that humans will attribute unsubstantiated powers to what should only be considered a useful tool. The section will conclude by declaring that it is likely that AI environments will develop using similar mechanisms that religions used.

The research will conclude in **Section 4** with a proposed roadmap for how AI environments could be integrated with humanity in a way that would encourage human individuality and also facilitate the self-determination of values, both of which are necessary features that need to be developed if it is to be the case that a broadly Kantian freedom is to be achieved. This solution will be proposed as a human-centric approach to a potential future where a significant amount of cognition is potentially outsourced to intelligent systems. This approach would ensure the AI environments avoided the dogmatic pitfalls that religions fell into during their integration with humanity. This will finally be contrasted with a less desirable potential future where AI environments continue on their current trajectory of resembling religions, in which humans are relegated to machines that propagate an ideology that is not founded on their best interests.

2 Origins of Religions and AI Environments.

2.1 Introduction and Approach.

Evolutionary biologist Bret Weinstein describes religions as 'compendiums of a kind of … non-literal wisdom' (Bret Weinstein on the Dawkins Debate 2018), and that we should perhaps not discard them merely as 'mind-viruses' as Dawkins' does (2016, 216). Dawkins' categorisation collapses the context of the phenomena into just being memes (ideas) that use humans in order to replicate. Weinstein insists that they are more like extended phenotypes and should be examined as extensions of the biological human, a sort of representation of our values, encoded in myths, texts, and rituals. This approach suggested by Weinstein is appropriate for examining early religious ritual, which owes much of its early success to leveraging biological traits and unconscious human needs. However, as religions evolve into complex social organisational systems, this examination through Darwinian evolution becomes overly complex. While their replication can technically be reduced to biology, these systems are better examined as memes because they become entities in and of themselves, much like Coca-Cola was once simply reliant on the taste-buds of humans to be successful, but is now a complex organisation that can own property and have a bank account, all technically describable through biological terms, but better explained through economics. The potency of what we recognise as religious organisations across humans from all cultures and generations means that an examination at the biological level would sacrifice cultural contextual detail, and is better explained through theories of cultural evolution than reduced to Darwinian thought.

2.2 Fundamental Origin of Religion: Unconscious thought becomes Myth.

Myths form the basis of most religions. They have similar characters and parables, many of which can be traced through time and across distinctly unique cultures. These common themes are referred to as archetypes. The psychologist Carl Jung was fascinated by archetypes and would collect artefacts symbolising common archetypes from varying cultures, mostly as a hobby, but with a strong belief that they could all eventually be linked

together through their common themes and examined empirically as having originated in the unconscious mind. One of his students, Erich Neumann, took this approach to studying archetypes. Jung praised him highly in a foreword to Neumann's book *The Origins and History of Consciousness* for having 'placed the concepts of analytical psychology … on a firm evolutionary basis, and erected upon this a comprehensive structure in which the empirical forms of thought find their rightful place' (2014, xiv). In this book Neumann traces the origins of archetypes back to unconscious human thought, as a means of understanding the base-reality that they originate from. This exploration provides clarity as to the reason there are so many common themes amongst myth, and subsequently; religion. Neumann's research suggests that myths are a representation of 'instincts' in the unconscious mind. '[archetypes and primordial images] are the pictorial forms of the instincts, for the unconscious reveals itself to the conscious mind in images which, as in dreams and fantasies, initiate the process of conscious reaction and assimilation' (Neumann, Jung, and Hull 2014, xv).

Jordan B. Peterson outlines in his book *Maps of Meaning* that the world can be understood in two ways, as a 'forum for action, or as [a] place of things' (1999, 1). By a 'forum for action', he means that the world can be understood in terms of values, which shape decisions and actions. He states that this 'manner of interpretation–more primordial, and less clearly understood–finds its expression in the arts or humanities, in ritual, drama, literature and mythology' (1). By this he is saying that these aspects of the world can probably be explained on an empirical basis through biological or psychological terms, but are best explained as a heuristic, or in his words as an 'interpretive schema that … guides action' (1). His book goes on to document how archetypes are the conscious manifestations of these unconscious 'guides', or as I will refer to them: the encodings of values.

This process marks the beginning of Weinstein's 'compendium': The conscious mind attempts to quantify and evaluate the instincts of the unconscious mind by creating archetypes and symbols (which later form into memorable stories, i.e. myths). In this process the conscious mind encodes unconscious knowledge by a form of aggregation and segmentation.

These archetypes and symbols maintain a strong intuitive resonance with the conscious

mind and survive (as a meme or as an extended phenotype) across varying generations and cultures. They may be modified relative to the context of the era but retain their intimate tether to the human psyche. This resonance means that these symbols can be used to connect with the human mind, potentially for means of manipulation, or as Dawkins would suggest: to install ‘mind-viruses’.

These archetypes should be understood as a de-centralised system of values. They are de-centralised because they originate spontaneously, they are not under conscious control, and outside of any infrastructure that can be intentionally constructed for the purpose of creating them. This makes their evolutionary success as an extended-phenotype or meme rather democratic, where ‘strong’ archetypes (in the evolutionary sense) gain an evolutionary advantage and ‘weak’ archetypes die off. This is not to say that one archetype is good if it survives, but rather that they constitute an encoding of a human value-system that is not consciously curated, regardless of intent.

AI environments have a similar method of encoding unconscious values, however this method does not rely on the human mind in order to process the unconscious into the conscious, but instead relies on statistical analytics and psychological modelling.

2.3 Fundamental Origins of AI: Raw Data becomes Profiling.

The core entity that enables an AI environment is raw data. A single piece of raw data is a quantified measure of an object, person, or event. Examples being things like shoe size, height, weight, bank balance, birthplace, voting history, or online shopping purchasing history. By and large, all raw data points are empirically observable measurements, quantifiable marks on a spectrum. With so much of modern life occurring on digital platforms, either professionally or personally, vast quantities of these raw data points are created, processed and stored every single day. This unprecedented creation of an entirely quantified and permanently recorded society has led to an explosion in information storage capabilities, and more significantly, information processing capabilities. This phenomenon of recording and processing vast quantities of seemingly insignificant data is called ‘Big Data analytics’, where the reduced cost to collect and store data has resulted in all quantifiable objects or events being collected without prejudice in its raw form, and later

evaluated for worth through different data mining techniques.
Typically, information (raw data) was evaluated ahead of time, deemed worth recording, and recorded only for the sake of routine operational tasks, such as book-keeping. These routine tasks categorised past events. Now we see these vast quantities of data being aggregated and processed in new ways to create predictions of the future, instead of recording the past. Multiple pieces of raw data can be combined to create a characteristic or trait, which is an aggregated calculation of smaller quantifications. Creating these categorisations, or profiles, of a collection of raw data relating to a past event facilitates a prediction of a future event, if the same pieces of raw data are available ahead of time. Using information to guess what might happen in the future is not a new phenomenon, however developments in the computational power, combined with the large stores of raw data continuously created and frequently updated by society, and also combined with the attention of some of the greatest mathematical minds to create predictive models, has resulted in techniques that can create incredibly accurate predictive models using seemingly unrelated data.

"Data was no longer regarded as static or stale, whose usefulness was finished once the purpose for which it was collected was achieved… Rather, data became a raw material of business, a vital economic input, used to create a new form of economic value" (Mayer-Schönberger and Cukier 2013, 5).

AI environments facilitate the collection of this data, the storage and transformation of it, and most importantly, the adaptive actions/choices it makes that feeds back to the user.

The data used in AI environments is generally collected across varying channels and interactions, a lot of which do not immediately appear to be related to the domain of interest. For example, an algorithm that makes predictions on political preference may collect raw data relating to arbitrary preferences of voters, such as what they 'like' on Facebook or the clothes they look at on-line. However, as we have discussed, data in its raw form is largely unusable due to the inability to separate the signal from the noise (i.e. to extract a discernible meaning from a collection of seemingly arbitrary measurements), and hence the need for the creation of characteristic profiles. Raw data generally does not have much variance between each line, it is all noise, no signal. i.e. All clothes sales generally look more-or-less the same, except a remote few where someone buys a large

quantity of items. In order to create a view of the data that contains noticeable variance, the data must be aggregated in some way. Perhaps the data could be aggregated by store: "stores with more square footage sell more bulkier items", or perhaps by region: "stores in the UK sell more umbrellas than stores in California", or by customer: "Customer X tends to buy neutral toned clothes, except for one brightly coloured item every Summer". Very simple character traits have been created here: big stores categorised = bulky items, UK stores = 'umbrella hotspot', customer x = 'neutral tones' and 'bright summer'.
By aggregating raw data, segmented profiles can be constructed: "all customers that have similar buying patterns to Customer X also happen to have similar average spends and ages, and tend to live in suburban areas (collected from loyalty card data)".

This principle of profiling can be harnessed at much greater levels than just purchasing history. Profiling is what drives the interactive and personalised digital media platforms. Online media platforms use profiling in order to show what they consider to be the most relevant (or typically, the most relatable) content to its users, in an attempt to increase the amount of time spent on the platform (in order to increase the amount of time they can charge for advertising). This has led to highly personalised news feeds which are infinitely scrollable, on which every hover, scroll and interaction is recorded for further personalisation. These records of user behaviour can also be used to make predictions outside of predicting what content the user may like to consume. Studies have begun to delve into the process of combining this behavioural data with psychological profiles.

A 2015 study found that creating a model of Facebook likes could predict behaviour of an individual better than any of the people most intimately close to them, including their own spouses, using only 300 interactions or less (Wylie 2019, 103).

These personality profiling models have developed significantly since 2015, and so have the ever-expanding corpora of human behavioural data captured by these digital platforms. Cambridge Analytica were the most recently significant and publicly examined analytics team to harvest huge quantities of personal behavioural data from social media and government records, model them, and using theories from psychology and sociology, accurately predict the unconscious personality profile of an individual (to later manipulate it for political motivations).

Combining unconscious behavioural data in this way creates a far more powerful profile of the citizen than a simple clothes-purchasing analysis tool used to sell more trousers, and infinitely more powerful than this data was in its raw state. This aggregation creates a profile with far more resonance with a person's unconscious mind, along with an increased ability to modify values outside of the original domain under which the data was created. This process is to create an *in silico* model of the unconscious values of an individual, or group of individuals, and to make them conscious by labelling and categorising them. Similar profiles are therefore categorised together in a process that bears a striking resemblance to the way archetypes are formed in relation to mythology.

This process shares a parallel with how archetypes are created as the conscious encoding of unconscious knowledge. The term 'conscious encoding' here must be carefully articulated. It is not, in either case, 'conscious' in the sense that it is deliberately shaped into this form. The conscious mind does not 'consciously' create an archetype, nor does a segmentation model 'consciously' create a profile. In both instances, the unconscious mind/model is merely taking raw data and shaping it into a coherent image that can be recognised by the conscious mind, which (ideally) captures the essence of the knowledge, separating the signal from the noise.

2.4 Are Big Data profiles accurate representations?

Those that used computational methods to digitise a human experience had an intimate understanding of the mechanisms at play and the full spectrum of experience that they were ignoring in the process of encoding discrete sections of it. These engineers or developers were generally just encoding a single linear mark on a spectrum of experience, aware that a full spectrum existed, but consciously encoding parts of it in order to digitise it. However, once these encodings begin to be incorporated into other processes, they begin to get 'locked in' (Lanier 2011, 7). An example of this, outlined by Lanier is the creation of MIDI, which constitutes a digital representation of musical notes. The MIDI standard is a standard of digital sound representation that encodes certain aspects of a musical note so that it can be replicated by a digital synthesiser. This encoding replicates the analogue sound at a single location on the musical spectrum. Due to the digital nature of this, it only allows them to either be turned on or off (i.e. it cannot replicate a half-

pressed key on a stringed piano, unless it creates 2 notes, one for full-press, and one for half-press, but even then it cuts out a quarter press and so on). This encoding greatly reduces the dimensions of sound that could be expressed (essentially turning an entire musical spectrum to a fixed set of sounds that could either be turned on or off). 'It could only describe the tile mosaic world of the keyboardist, not the watercolor world of the violin.' (2011, 7). At the time of creation, this was not an issue because it was only *intended* to replicate a narrow dimension of the musical spectrum in order to interact with a synthesizer. However, MIDI became a standardised way of representing sound in digital software, and eventually so many products came to depend on it that it became engrained as the standard for musical expression in software. This is what Lanier refers to a getting 'locked in'.

This is acceptable practice in technology, as typically the consequences of the lost sections are not monumental, however issues arise when these encodings are used in order to build a profile of a human-being, because these quantifications are only a general mark on their spectrum of experience. The example of MIDI is to illustrate the process, rather than to be an example of all of the ways experience will be 'locked in'. The process to be identified is the process of iterative distillations of raw experience. This has significant implications in AI environments because they cycle through these distillations much faster. The musical spectrum was only distilled once in the case of MIDI. Whereas AI environments consistently distil profiles of individuals, typically in the pursuit of optimistation and attempting to find the perfect measure to quantify a person for whatever narrow purpose the model is being used for. For example, take a person's voting history. Do these single marks adequately represent the person's entire political view? They undoubtably provide an indication toward their preferred candidates, but they do not fully represent the spectrum of their political beliefs. Lesser still the profile would be generated by this data (e.g. 70% party A, 15% party B, 15% unknown). As will be explored later in this thesis, the issue is further intensified once they are passed to an artificially intelligent decision-making system, where it uses these inaccurate representations to make decisions on behalf of the person. This 'lock in' of encoding (the way that political votes are represented in this instance) is entirely out of the person's control, and yet could have significant implications for decisions made on their behalf. The incremental loss of experience at each stage of encoding, iterated over many cycles, could result it greatly reduced conceptions of defining

traits. This process of reduction is often understood as distillation and optimisation, and interpreted as a good thing, however realistically it results in statistical models that are not able to generalise. This inability to generalise means that a model can become overfitted (too entrenched in its ways, unable to come up with novel solutions), and therefore biased. If this system is making predictions, then they will be inaccurate. If the system is making *prescriptions*, then they will be ever gradually converging toward a mono-culture (i.e. constantly trimming the edges of experience). The societal implications of which will be discussed in Section 4.

Much like the origin of archetypes, the initial creators of these encodings had an intimate connection to the creation process. While the 'creators' of archetypes may not have had as much empirical understanding of the origins of their encoding, they were at least only one step removed from the process, in which they alone stood the best chance of reverse-engineering through introspection, and retained full control over how their 'program' was used. As we shall see later, as further additions, intermediary parties, and dependencies are introduced, the original creators of these programs become less able to maintain control over the ways in which they are used.

2.5 Conclusion

As archetypes became the foundational principles of myth, Big Data profiles are the foundational principles of AI environments.

Archetypes that originate from the unconscious mind ensure an intimate subliminal link is retained between the agent and the idea. The idea may conceal itself in various guises, as we see with different myths across different cultures that retain consistent archetypes encoded in them. The embedded conscious manifestation of unconscious values means that these stories appear recognisable, relatable, and trustworthy to the agent. This unconscious encoding could facilitate the illusion of the superfluous guise being wrongfully deemed a trusted source of information. This can facilitate a third-party agent using this 'trusted source' in order to adjust and shape values of those that feel connected to the stories (the archetypes) that are used.

Modern digital advertising techniques use similar methods in order to influence opinions of their targeted audience. By creating a digital encoding of a person's implicit value-system, these agents can manipulate the content that the person sees until targeted aspects of their value-system are either exaggerated or eradicated. This can be used to sway public opinion at the level of a community, a region, or even a nation, as was recently the case where Cambridge Analytically effectively used Big Data analytics, behaviour psychology and digital advertising to disrupt and sway political campaigns (Wylie 2019).

Outside of the realm of nefarious actors, there is an issue in which these encodings sacrifice the full spectrum of experience in the process of recording sections of it. While these are still encodings of unconscious behavioural data, it should be noted that they are just that: encodings. At this stage in the research, this issue does not have an immediate impact on the comparison between religions and AI environments, however we will see this issue emerge as a considerable concern when evaluating the legitimacy of decisions generated by AI environments.

In conclusion, what is currently relevant is that these archetypes and profiles allow mass communication and connection across large groups of humans by isolating unconscious common values. In this way, other values can be shaped through behavioural modification. This process becomes very susceptible to 'pretend priests' because of its intricate link to the unconscious mind, and resistant to investigation through its apparent complexity.

3 Evolution of Religious concepts.

3.1 Introduction.

This section will explore how the common value-systems discussed in the previous section moved from being decentralised archetypal values to being curated centralised collections of myths, which go on to become religious concepts. This process begins with the introduction of intermediary parties and evolutionary mutations and adaptations in ritualised behaviour. I will examine how certain biological traits impacted the evolutionary fitness of religious concepts, and what necessary features these concepts all share. In this section I will evaluate the biological and cognitive mechanisms that were instrumental in amplifying these early religious concepts, and how they were propagated through ritualised behaviour.

Later, religious concepts will be explored as memes, departing from the extended-phenotype classification they had been examined under. This new mode of evaluation will allow religious concepts to be evaluated as evolutionary objects in their own sense, rather than just an extension of the human biological structure. In this way, it will be possible to observe how these religious ideas interact with the mind from the outside, in order to document the behaviours and traits they encouraged or discouraged.

The final evaluation of this section will consist of a comparison between religious concepts and similar themes observed in AI environments, under specific headings relating to the way that they interact with the human mind. This comparison will attempt to comprehensively determine whether the human mind has the capacity and proclivity to interact with AI environments in the same way as they would with a religion, as outlined as the objective of this research.

3.2 Myth as a Societal Function.

“Nothing is surely more intangible and unreal than fictions, illusions and opinions; and yet nothing is more effective in the psychic and even the psychophysical realm.” (Jung, Dell, and Baynes 2001, 229).

Communal adoption of mythology became a way for members of a society to align themselves toward a common value-system. This view of mythology as a 'charter for social action' is referred to as 'functionalism' (it should be noted that the term 'functionalism' is also used to describe areas of thought outside of this specific context, which are not necessarily related). Functionalism maintains the view that 'myths are narratives formative or reflective of social order or values within a culture' (Magoulick 2004).

The archetypes contained within these myths ensured that they remained intimately linked to the human psyche, and the narratives of the myth provided abstract advice on the highest good to be achieved, or the lowest evil to be avoided. This meant that not only did they resonate with their audience, making them more likely to be replicated, but at the same time also actively shaped their audience's internal morality toward a 'common network of stories' (Harari 2016, 170). This allowed for harmony and co-operation in large groups of humans, where the common value-system was continually reinforced by every other member of the community. It is important to note (for further discussion later in this research) that prior to any formal grouping of these common networks of stories into religions, that they could be understood as a type of decentralised universal value-system, largely analogous to early speculation of what the internet would be, as a self-regulating decentralised hive-mind of human wisdom (Lanier 2011).

These common networks of stories continued across generations, and gradually evolved to include rituals, ceremonies, songs, physical artefacts, and traditions.

3.3 Introducing intermediaries : Biological roots.

From an evolutionary fitness point of view, by attracting invested owners, religious ideas found a way to become stronger, but at some costs. A cost of acquiring an owner is that some responsibilities are outsourced. Take, for instance, a sheep. While ultimately in the best evolutionary interests of a sheep, breeding and food choices are outsourced to the shepherd. This is both convenient and inconvenient for the sheep. Conveniently the sheep no longer needs to take up any more of its time with making these decisions for itself, essentially it can now tag 'along for the ride', so to speak. Inconveniently, however, this

reduces the sheep's agency, which may not sit well with a sheep that has very strong independent values (not what we generally consider a typical attitude for a sheep, but perhaps once upon a time they were highly independent animals). This amounts to a divide being created between the agent (i.e. the sheep) and its agency (i.e. the choice to eat the grass in the lushest field, three fields away from where it is currently grazing). This agreement (of sorts) can be of mutual benefit to both parties: the shepherd earns a living, and the sheep leads a better life (quantified by reproduction rates and average lifespan). However, there is a potential for a backfiring in this mutual agreement, in so far as the sheep may forget how to be a sheep. Or should we say: the original core qualities that the species of sheep possesses may change based on modifications and evolutionary selections regarding newly beneficial characteristics in new environments. Repeated over enough generations of wild sheep, the shepherd transforms a wild animal that can reproduce, feed and live independently into a dependent commodity. This concept, applied in the context of this research and fully developed ultimately asks the question: does outsourcing in this way create an inability to perform that function independently, if maintained over generations? This exploration will begin by examining the first intermediaries that came between 'the answers' and those that created them, and what the initial biological benefits were for introducing them, like the sheep that reproduces more and that lives longer.

Anthropologist Pascal Boyer explores many facets of human behaviour, with special regard to religions and ritualization. In a paper with psychologist Pierre Liénard, the authors examine why humans have a tendency to ritualise behaviour and propose that 'the occurrence of ritualization depends on the conjunction of two specialized cognitive systems… a motivational system geared to the detection of and reaction to particular *potential* threats to fitness… [and an] other system [that] might be called "Action Parsing"… concerned with the division of the flow of behaviour into meaningful units' (Boyer and Liénard 2006, 595).

The authors evaluate these cognitive systems at an individual level as being most prominently activated in subjects with certain disorders (such as Obsessive Compulsive Disorder), or during childhood development (ages 2 to mid-childhood), or particular phases in lifetime (parenthood). However they also evaluate these cognitive systems being

activated at a cultural level, where social occasions can motivate such ritualistic behaviours.

Boyer and Liénard state that they ‘consider that cultural rituals may be better explained … as partly parasitic on the Hazard-detection and Precaution systems’. This speculates that all humans have the capacity, or even maybe so far as the inclination, for ritualising behaviour, across the individual and societal level. This occurs most prominently while experiencing anxiety during the detection of potential danger and fitness threats (609).

In the same paper, the authors also discover that rituals share some common features, such as:

- Compulsion: ‘there is an emotional drive to perform the action, often associated with some anxiety at the thought of not performing it’.
- Rigidity, adherence to script: ‘strive to achieve a performance that matches their representation of past performances and attach negative emotion to any deviation from that remembered pattern’.
- Restricted range of themes: ‘Many rituals seem to focus around such themes as: pollution and purification, danger and protection,… the construction of an ordered environment’.
- Goal-demotion: ‘Rituals generally include action-sequences selected form ordinary goal-directed behavior. But the context in which they are performed… [is] divorced from observable goals’.

We will return to these features later when making comparisons with AI environments.

This research can help to explore and understand the cultural transition from mythological stories to ritualised behaviour. Mythology facilitated a general social calibration/orientation toward a common system of values, however the ritualization of behaviour allowed for increased health, which accelerated individual replication and societal progress as well as having an impact on cultural attitudes toward authority and ritual.

3.4 Introducing intermediaries: Biology and the role of Ritual.

3.4.1 Leveraging Biological Evolution

Shamans were an early example of intermediaries that emerged and contributed towards a divide between ‘the answers’ and those that sought them. Instead of turning inward (ironically to the true source of ‘the answers’) the population began to turn to an intermediary authority figure that maintained the link to ‘the answers’.

Mythology, when interpreted literally, implies a duality of material and spirit. It implies the existence of a separate world filled with spirits, Gods, or essences of good and evil. While unaware that the mythology was the natural result of their unconscious minds, societies looked to a shaman in order to bridge the gap between them and the world of spirits and essences. The shaman played various roles in the society, such as a communicator with the mythical and spiritual world, a healer, and a curator of tradition. The vast majority of these roles included some sort of ritualistic behaviour. As mentioned, the shaman was a healer to the tribe or society they were a part of, they healed their patients by appealing to the spirit world, or through the use of natural remedies. Frequently the two were combined together, reaffirming the dualistic belief held by the people of the time.

Shamanic ritual was an early form of ritualised behaviour that improved mental and physical health through ritualistic acts (as well as some natural remedies, most of which were not medically understood, but possessed healing properties none-the-less). As we will see, the belief in the authority and legitimacy of the shaman’s ritualised behaviour shifted the patient’s cognitive state from the heightened potential-danger-detection state into a satiated healing state.

The immune system of an injured animal will not attempt a costly healing procedure until they are within the safe confines of their den or nest. This is because it must reduce the amount of energy being dedicated to major organ functions, which would most definitely

be unwise in an unsafe environment. In the case of humans, a similar mental reassurance is required in order for the immune system to dedicate the resources required to combat an affliction. As suggested by Boyer and Liénard's work, a human will seek ritual more prominently when in this state of 'potential danger detection'. It is the belief in ritualised behaviour that can make a human know they are safe, at which point their body can enter the satiated mental state required for self-healing.

As an example from personal experience; a colleague of mine who is currently pregnant recently had some abdominal pain. She reported that the pain was quite severe and she had decided to call her doctor for advice. The doctor reassured her that the pain was not anything to be worried by, but rather quite normal, considering that her stomach was going through a process of continuous stretching. The doctor told her that if the pain was too much that she could take pain killers, at which point my colleague noticed that the pain had all but dissipated. She recalled that it was as if her body had exaggerated the pain in order to draw her attention to it (or at least hadn't used any adrenaline to mask it), until she had consulted with an authority figure, at which point the body was reassured that the pain was normal and proceeded with natural forms of pain relief. There is contemporary neurological evidence of this mind-body connection, examples of which are the use of mindfulness meditation as an effective means of combating chronic pain (St. Marie and Talebkhah 2018), as well as the well-documented placebo effect (Tétreault et al. 2016).

This common trait of the mind made shamanic rituals rather successful. The only condition in the success of the ritual would be the patient's belief in the legitimacy of the act. In an attempt to create a similar belief in legitimacy, modern doctors display their certifications on the wall of their office, not in an attempt to boast their achievements, but rather to display that they are highly qualified, and should be trusted (Cohn 2019, 109).

By leveraging the human tendency to seek ritualised behaviour during times of distress, and through the immune system's capacity for self-healing (combined with some genuine natural remedies) shamanic ritual attracted patients and made the populous healthier, primarily with regard to members with strong belief in the validity of the ritual. Speaking from an evolutionary perspective, these healthier member of society are more likely to reproduce in stronger numbers. Due to this occurring over the course of generations, it is speculated that the introduction of shamanic ritual healing (and other ritualised behaviour)

contributed toward an intensification within the gene-pool of the propensity to believe in authority and ritual practices (Dennett 2007, 158).

3.4.2 Ritualised behaviour

A comparative phenomenon of ritualised behaviour is explored in the book *Re-engineering Humanity* by Brett Frischmann and Evan Selinger (2018). Throughout the book, the authors explore phenomena surrounding the ritualization and automation of human behaviour in digital environments. They highlight some of these processes as 'creeps', using this word to describe processes that are generally micro-transactions (i.e. agreeing to the terms and conditions of a website without actually reading them) that typically have genuine reasons for existing, but that over continued interaction gradually habituates the user to a state of passive compliance, which has a much larger aggregated ramification than any of the constituent actions.

Of two types of creep that I will address in this section, the first creep is 'electronic contracting creep' (71). Contract creep comprises of users that become so habituated to lengthy electronic contracts, where 'human deliberation…tends to be unproductive' (72). Through this process, the users become habituated to accepting a legally binding contract without any deliberation whatsoever, because it becomes normal operating practice when they visit any new webpage or app. This is partly by design: Electronic contracts are designed to 'minimize transaction costs' (i.e. be user-friendly), 'maximise retention' (i.e. don't drive the user away from the site), 'minimize design and operational costs' (i.e. simple as possible: display lots of contract jargon and an 'I accept' button), among other objectives. These constitute 'non-negotiable, incomprehensible, and seemingly endless contract terms [that] can be so daunting that no one bothers. This design feature affects consumer behaviour on a micro contract-by-contract basis' (74). Through this practice, we see perfectly rational decisions on behalf of the user (accept a contract this is *probably* not worth reading) accumulating into an intensification to trust potentially untrustworthy authority figures.

The second creep that is addressed earlier in the book is 'Surveillance Creep' which refers to the habituation 'to submitting data to opaque third-parties that exercise authority' (it

should be noted that in this instance the authors were speaking specifically about students encouraged to submit biometric data to a third-party platform, but I am extending the scope of the quote to capture the broader definition). Through micro-transactions from a young age this type of creep habituates the user to submitting their personal data without fully evaluating the party it is being submitted to, and combined with contracting creep, can lead to legally voluntary exploitation or manipulation.

Contracting creep highlights the capacity for an initial passivity in the face of complexity resulting in a gradual habituation to trusting intermediary authority figures. Surveillance creep highlights the capacity for a trust in intermediary authority figures leading to a willing submission of private (and valuable) information to unknown intermediary parties wrongfully thought to be the status quo of using digital platforms. These types of creep become ritualised behaviour, where we see users willingly submitting data to intermediaries under the assumption that it is the only way to continue to avail of the digital platform and to continue the normal functioning of society.

In addition to the explanations for ritualised behaviour at an individual level, this behaviour also plays a significant role at a societal level too. Pascal Boyer provides an anthropological explanation for public ceremonial rituals such as weddings, baptisms, and funerals. His explanation is that these ceremonies act as a means to announce 'that people's interaction will indeed be recalibrated to accommodate parental investment' (2001, 247). In this instance he is referencing specifically announcing the birth of a child, which in some religious contexts will not be considered as *actually* born until it is ceremonially baptised. In essence, this public ceremony acts as a means of announcing that your social contract is changing, as your priorities change, and that your community and religion can expect your behaviour to change in the near future. It seems perfectly reasonable to speculate that this allows for continued harmony amongst the community, where undue responsibility is not expected of participants that have publicly announced their inability to accept further responsibility outside of their immediate kinship, at least for the time-being. We can further speculate that this helps avoid issues that can cause internal conflicts amongst groups, such as the 'free-rider problem', as outlined by Rodney

Stark and Roger Finke, who in their book *Acts of Faith* explore religion in terms of economic theories and rational choice (Stark and Finke 2007).

An early version of a similar phenomenon can be seen in AI environments right now, and we can also speculate as to how these features may benefit the development of certain platforms over others. There is a scene in the movie *The Social Network* (2010)–a movie that dramatizes the inception and early developments of Facebook– that sufficiently encapsulates the concept of updating social contracts in digital environments. In this particular scene, Mark Zuckerberg (founder of Facebook) is close to releasing the first version of Facebook, however he feels that there is still a critical feature missing. In the scene, a friend is telling Zuckerberg about a girl he likes, and makes reference to wishing there was a way to find out if she had a boyfriend or not. Zuckerberg darts back to his student dorm and begins coding the final feature of his platform before putting it live: the relationship status. It is easy to conceive that an AI environment that encourages the continual update of social contracts (through relationship statuses, digital wedding invitations, photo-albums) maintains a replicative advantage over one that does not. Not only that, but that they also receive the indirect benefit of having contextual data on that person continually updated. It is likely that these major life events are significantly influential in behavioural modelling and prediction. For example, a mid-twenties male that has just left a relationship that is looking at sky-diving videos is probably a lot more likely to buy tickets if nudged in that direction, as opposed to a mid-fifties married male with 3 children watching the same videos. This tendency to publicly announce significant life events can prove very beneficial to AI environments seeking to model human behaviour, and will continue to be encouraged by these environments based on evolutionary advantages.

3.4.3 Evaluation of Ritualised Behaviour in AI environments

We can observe ritualised behaviour in humans interacting with AI environments that span across the 4 key themes previously introduced by Boyer and Liénard. Those themes being: compulsion, rigidity, restricted ranges of themes, and goal-demotion.

- Compulsion: ‘there is an emotional drive to perform the action, often associated with some anxiety at the thought of not performing it’.
    - The majority of interactions with AI environments happens through smart mobile phones, as the primary technology that people use to access digital platforms (Global Digital Overview, 2020). These mobile-first platforms facilitate the interactions that create behavioural datasets and also facilitate the subsequent exposure to behavioural modification. Recent studies indicate increased levels of anxiety in test subjects when separated from their smart-phone (called ‘Nomophobia’), and a compulsion to interact with their phone even when they are aware there are no new notifications (Rodríguez-García et al., 2020; Kuss, Griffiths., 2017). Social media addiction is also a compulsive action, that has been make further difficult to satiate following the introduction of the ‘infinite scroll’ news feed (Montag et al, 2019, 4).

      These compulsive rituals contribute to the continued creation of behaviour datasets on individuals, and the continuous exposure to the behavioural modification material that the environment broadcasts.
- Rigidity, adherence to script: ‘strive to achieve a performance that matches their representation of past performances and attach negative emotion to any deviation from that remembered pattern’.
    - The rituals that people create with interacting and contributing to AI environments through their actions and behaviours are recorded in databases. In order for behaviour to be accurately recorded it must be quantifiable and for that to be the case it must be constrained in some way. The data must be given a lower and upper limit on size, adhere to particular data types (integer, float, string, date) and given these constraints, most ways of interacting with AI environments are fixed with rigid rules around acceptable behaviour (not necessarily moderation, but with regards to creative format). For example, Twitter has a character limit on its tweets, most Instagram posts consist of an image accompanied with some text and hashtags. This has become easier with the introduction of aggregated platforms, such as Facebook, which contains a marketplace for businesses

to make fixed profiles instead of building their own website (that would be hard to quantify and track) as well as individual users that can arrange their profile based on a customisable –but heavily constrained– layout. There is a competitive advantage to a platform that encourages its users to adhere to rigid scripts, because it means that they receive the cleanest datasets. This is especially true for creating behavioural datasets of users: for example, most social media feeds only allow one dimension of movement (up or down), 3 actions (like, comment, share), and one option to contribute (add media). This makes behaviour tracking very easy for the platform, where it is technologically trivial to track how long someone looked at a piece of content (generally only one piece of content will fit on the screen at one time, sometimes the user will be able to see a portion of the next piece), whether they clicked on it, liked it, commented on it, shared it, or decided to add their own content. This used to be extremely difficult for platforms to do, especially where desktop sites are concerned, because that would require eyesight and mouse-hover tracking, which affected site performance. This new, minimal, constrained model makes for very simple, consistent and clean datasets that can be mined for behavioural patterns. This layout also means that advertising content is easily added into the feed, and easy to quantify in terms of behavioural impact.

- With regard to the 'negative emotion' in deviation from the rigid pattern, we can see evidence of public discomfort following any major changes to the user interface (UI) of these platforms. For this reason, these platforms general release changes in sporadic and minor ways, or hold out making major overhauls of their UI unless absolutely necessary. In evolutionary terms, the benefit of this rigidity in the platform is that it preserves fidelity of transmission (behavioural data is high fidelity and comparable across different platforms and across time) and also it avoids evolutionary mutations that might make the system less likely to reproduce (in the case of AI environments, less likely to have users interact with it).

- Restricted range of themes: ‘Many rituals seem to focus around such themes as: pollution and purification, danger and protection,… the construction of an ordered environment’.
    - The concept of online profiling and personalised content has led to a vested interest in avoiding pollution and encourage purity on behalf of the person being profiled. It is now commonplace to actively maintain an updated profile across social, professional, and personal domains.
    - This can be motivated through social obligations such as relationships (to not make a relationship public implies that it is something to be embarrassed of or that it is being kept secret), actions like showing likes or dislikes for a friend’s new post or local business.
    - There are also professional obligations to keep profiles updated, such as a job-change (with online profiles acting as digital CVs, then they must be kept up-to-date, you can no longer state publicly that you are a representative of a company you are no longer employed by).
    - Additionally there are personal rituals motivated by feeling overwhelmed if there are too many unread notifications on a device (and the subsequent relief when they have all been cleared), or a hesitation to let another person use your profile on a platform out of fear that they will pollute your feed as a result of the algorithm learning their preferences instead. This is complemented by people also taking proactive approaches to purifying the algorithm to accurately represent what they like by providing lists of likes and dislikes when signing up, or providing ratings and feedback that are only intended for the algorithm to see.
- Goal-demotion: ‘Rituals generally include action-sequences selected from ordinary goal-directed behavior. But the context in which they are performed… [is] divorced from observable goals’.
    - Boyer provides some explanation for this detachment of actions and observable goals in ritual behaviour, when explaining why it is difficult to find and compare similar rituals across religions because ‘Evolution does not create specific behaviours; it creates mental organisation that makes people behave in particular ways.’ (2001, 234). Goal-demotion is a difficult

theme to assess, due to some differing comparisons between religion and technology. In religion, it is very common for the goal of a certain action to be completely unobservable to someone who does not know of the religious reason for the action. For example, imagine you are observing a shaman burning tobacco leaves in-front of some statues. Without any insight into the internal motivations of the shaman, or knowledge of their beliefs, the goal of the shaman is not apparent (unless, of course, the goal is to make the statues smell like burnt tobacco leaves, which we can assume it is not). With technology however, there generally is not their air of metaphor to the rituals as is seen in religions. One can suppose that supernatural beings can only be communicated with in unconventional ways (even spoken prayer has some form of unintelligible ritual. Would you kneel and close your eyes in order to speak to your next door neighbour, or even a celebrity you greatly admired?). Whereas communication with AI environments (the un-supernatural-ness of which we examine later in the research) is quite achievable through normal means of communication. However, this raises an interesting point. Our communication with AI environments makes sense because we all understand the behaviour. The shaman's behaviour does not make sense because it is not behaviour that is intuitive, but that does not mean it is any less valid. Technologist and AI pioneer Ben Goertzel frequently uses this example in interviews when questioned about super-intelligent AI and what it will look like: Goertzel says that his dogs must think that he is very strange every time they see him sitting at a computer, because in their ontology, he is probably guarding this box for 8 hours of the day, occasionally taking a break for food, and sometimes pressing it. It isn't in their understanding of reality to think that he is actually programming an algorithm on multiple servers in China, or communicating with another researcher (Fridman 2020). Essentially, any interaction with AI environments will look like goal-demotion to the uninitiated. In the exact same way that religious rituals are completely unintelligible to the uninitiated, but to the initiated: the action is the complete opposite of goal-demotion, because of course, the only conceivable goal of burning tobacco

> leaves in-front of a row of statuettes is to 'cure someone whose mind is held hostage by invisible spirits' (Boyer 2001, 1), what else could it possibly be for?

This ritualization of behaviour helps the transition from a collection of features to creating the infrastructure for full-scale AI environments, where every aspect of everyday life is mediated through it. Now that the infrastructure has been built (i.e. the rituals are in place), we can examine how the human mind propagates it (AI/religion as a meme).

3.5 Religious Evolution: Religions as memes.

It is at this point that we migrate away from Weinstein's view of religion as an extended phenotype. As has been previously stated, it makes sense to interpret early religious phenomena through the lens of biological evolution, however eventually it is more practical to interpret religious concepts as entities themselves capable of replication (whilst of course not forgetting that they continue to rest upon—and can be influenced by—biological evolution).

From this point onwards, we will evaluate religious concepts as independent, and examine ways that they are replicated, the process of which is referred to as 'memetic evolution' (Dawkins 2016, 228) or 'cultural transmission' (Dennett 2007, 78). Like biological evolution, ideas (also referred to as 'memes') need to be good replicators in order to survive. An idea in the heads of two people is more likely to survive and be passed on, than the idea in the brain of one person. Unless, however, that one person is twice as 'fit' as the other two and more likely to survive. Cultural evolution can be weighted and influenced by biological evolution and vice versa (e.g. a culture of abstinence in every member only lasts one biological generation). Often, it is in the meme's best interest to make their human hosts *biologically* fitter. Or better still, to manifest themselves outside of the human host, as an object (e.g. cave painting, monument, book).

Up until this point, biological traits have made religious ideas very strong replicators, with the human mind being primed to retain and replicate them in other minds. Cultural evolution, however, is what facilitated an accelerated spread of these ideas, as cultural evolution allows more iterations per generation, which is extremely beneficial for increased mutations and replications. However, this also means that religious ideas potentially become subject to too many mutations which could dilute some original evolutionary strengths. Therefore, the strongest ideas are those that are replicated quickly but also with high fidelity. This can be facilitated using agents (such as shamans) that maintain the integrity of the idea across transmissions, or it could be a trait of the concept itself that insists on perfect fidelity (either through prescribed ritual, or by strict rules of adherence).

By exploring religious evolution as memes (i.e. as ideas), there is an implicit statement made that, from here on, religions will be considered as the result of cognitive processes. This is indeed the epistemological approach that shall be adopted from this point forward. This approach of examining the cognitive science of religion has been criticised for not actually explaining religions, but rather explaining the 'mental mechanisms without which there probably would not be religion… we are not so much on the way 'Towards a new science of religion' as we are in the realm of psychological explanation' (Sinding Jensen 2009, 130). The central issue that this particular author has with the cognitive science of religion is that "if the occupation with religion as subject matter is in fact applied to say something about our evolved 'mental architecture', then the 'cognitive science of religion is not so much about religion as it is about how humans interact with imagined agents' (2009, 145). In essence, the author's concern is that mental mechanisms alone cannot sufficiently account for all religions from origin to current day because minds only *process* religions, that they are not *actually* religions. For example, like trying to understand a computer programme by observing a CPU. These are indeed valid reasons that the cognitive science of religion is not a sufficient approach to explain all facets of the phenomena, however for the purposes of this research, it is sufficient to observe the cultural evolution of religious concepts. The following section will explore the cultural evolution of religious concepts, each method of which relies on mental mechanisms in

order to replicate. Ultimately this research is concerned with how human minds are primed to replicate collections of features, of which religions have been one and AI environments proposedly another. The purpose of this section is to document the mental mechanisms that allowed religions transform from simple biological by-products to large-scale organisations of which almost every human in contemporary history was a part of. It is not so much concerned with what the programme does, but rather why it passed through this particular CPU more effectively than another.

Memetic evolution is *complementary* to biological evolution. As Daniel Dennett specifies 'replicators are dependent … on replicative machinery that is built and maintained directly or indirectly by the parent process of biological evolution' (2007, 344).

A strong indicator of this is that similar fundamental archetypes can be found across cultures (as previously explored in the works of Jung, Peterson, and Neumann), so the individual culture cannot be the instigator of these memes, they must contain some types of common cognitive traits that are shared by all minds.

Dawkins puts forward theories of how a brain developing the tendency to listen and retain knowledge from authority figures most likely enhanced the chances of survival, from a biological point of view. But in a subsection titled 'Psychologically Primed for Religion' he also extends this theory to include how these developed features (or 'modules', as he calls them) of the brain also potentially had some side-effects that may have been of benefit to religious concepts themselves (2007, 209). These 'mis-firing' 'modules' could act as trojan horses for humans to adopt beliefs that are not necessarily in their best interest.

'The religious behaviour may be a misfiring, an unfortunate by-product of an underlying psychological propensity which in other circumstances is, or once was, useful' (Dawkins 2016, 202).

In this case, he draws an analogy between religious behaviour and the way in which a moth is drawn to light (and subsequently burns to death in the candle). The traits that enabled a species at one point (in the moth's case: a reliable way to navigate prior to the invention of

artificial light), given further developments elsewhere, may have adverse and unintended effects.

In much the same way, Dan Dennett also outlines in *Breaking The Spell* that we have quite a few 'features of our minds' that 'sometimes have curious by-products'. Most notably, these features are 'sometimes ripe for exploitation by other replicators'. These features 'interact with one another in mutually reinforcing ways, creating patterns observable in all cultures' (2007, 107). In this context, the 'pattern' being referred to is what we would consider a religion. These patterns being made up of collections of ideas (values) about the world which are adopted by a human host.

Pascal Boyer refers to these patterns as 'concepts'. Boyer highlights that all of these religious 'concepts' share four necessary conditions, even across cultures.

Religious concepts *must*:

- Make recall and communication easy
- Trigger emotional programs
- Connect to our social mind
- Become plausible and direct behaviour

(Dennett 2007, 107).

Over the following section, we will build a collection of features of religions that we can evaluate using these necessary conditions and compare to similar mechanisms that appear in AI environments. This comparison will indicate whether AI environments have the potential to resemble religious concepts. If these AI patterns satisfy these conditions then we can infer that they resemble religious concepts, that they will follow similar evolutionary paths, and will be able to leverage similar evolutionary 'misfires'.

3.5.1 Psychologically primed: Hyperactive Agent Detection Device (HADD):

Nature has developed the underlying tendency to outsource to reduce costs, but there is still the cost of calculating whether it is reasonable to outsource to an entity. For example, why will a person outsource a decision to a digital personal assistant ("Alexa, who should I

vote for?"), instead of a coin flip ("Heads: Candidate A. Tails: Candidate B")? It is important to note that in this example, both Alexa and the coin are acting only as tools that facilitate the connection between the person asking the question and the decision-making entity (like the Shaman did in the previous examples). Alexa is the link to the knowledge contained in whatever search engine it is linked to, and the coin is the link to the knowledge 'the controller' has (God, deceased relative, fate etc). The term 'belief' seems to be an appropriate proxy measure for this concept of 'calculating the cost of whether or not to outsource'. A person either believes in the entity controlling the outcome, or they do not. A contributing factor to this belief is the idea that the entity has a broader view or deeper knowledge of the scenario than that of the questioner. In religions, this takes the form of the questioner projecting omniscience onto Gods, Saints, or deceased relatives (Dennett 2007, 125–31). In digital forms, this omniscience is reflected in a trust in personalisation optimisation and information retrieval. This belief in an omniscient agent should be built up at a the level of personal interactions in order to be most effective (Leeuwen and Elk 2019).

The Hyperactive Agent Detection Device (HADD) is a cognitive mis-fire that attributes agency to items or events, manifesting as an intuition that the item or event is being controlled. This hyperactivity is likely to have evolved in quite mundane contexts, where there was an evolutionary advantage to believing that a rustle in a bush was a tiger (even when it was not) rather than believing it was something harmless (even when it was a tiger). The cost of these false positives (the false belief that there is a tiger in the bush) is neglible (perhaps a slight outlay of adrenaline and a few more calories burnt than necessary), especially when considered with the cost of a false negative (the false belief that there is nothing in the bush, when in fact there is a tiger). This leads to an evolutionary advantage in hyperactively attributing agency.

The HADD creates the illusion of a personal interaction within the arbitrary event that occurs, this detection of a personal level interaction then encourages and amplifies a belief in a potentially non-existent controlling agent. In religious circumstances, this is 'God answering your prayers' or the Holy Mary presenting herself to you. It can manifest in very physically affirmative ways such as the Holy Spirit entering your body (Evans 2018, 31). In AI environments, this phenomenon can be seen in recommendation engines and

dynamic advertising. Often an end-user will see a post, such as an advert for a skiing holiday. This person will be shocked or surprised by the fact that they had recently been thinking or talking about going on a skiing holiday and now the algorithm 'knows'. In reality, it is either because they went to a travel website and now have a cookie, or their phone is recording their conversations, or perhaps they are just subjects of their demographic profile who typically book skiing holidays at this time of year. In this particular instance, the reason the advert was presented to the person is irrelevant, but what is important is that this 'interaction from the Gods' is re-affirming the necessary personal aspect of the human/AI interaction, which then encourages the projection of omniscience from the human onto the entity. The person will likely turn to their algorithmically driven companion the next time they are considering a getaway break. We humans seek to do this because once that relationship is made then that entity can be used as a reliable agent to outsource a mental process to. Each time this occurs, the person replicates a meme of AI environments possessing some sort of over-exaggerated inexplicable complexity, often distilled to a simple attribute of omniscience.

3.5.2 Psychologically primed: Are algorithms the new supernatural?

Further to the tendency to project agency onto phenomena that lack it, the human mind is also more prone to recalling the instances when this projection happens, which further intensifies the recursive loop of agent detection and intelligence projection [reframe: spirits aren't always intelligent. Sometimes they are just powerful but easily tricked].

In his book *Religion Explained*, Pascal Boyer explores the necessary conditions that give rise to supernatural features in religious concepts. He states that these representations of supernatural agents are the 'particular combinations of mental representations that satisfy two conditions.'

1. 'The religious concepts *violate* certain expectations from ontological categories.'.
2. 'They *preserve* other expectations.' (Boyer 2001, 62).

For example, a religious concept that satisfies these conditions is: there is a person that is invisible, who cares about you and listens to your problems. This concept violates the expectation of the ontological category of a person (i.e. a person cannot be invisible), but

preserves the other expectations of the ontological category (i.e. a person can be expected to exist, care about others, and listen to problems). However, the supernatural concept cannot violate all ontological expectations, as the referential framework becomes unrecognisable and unmappable in the broader ontology. All religious concepts that include some form of the supernatural share these features.

Boyer found that concepts that have this feature are also more likely to be recalled, even in long-term studies (over the course of months). Cases where there is a 'violation of ontological expectations' is recalled more effectively than one with just 'mere oddities' (2001, 80). The example he uses is a concept of 'a man who can walk through walls' is recalled more effectively than 'a man with 6 fingers'. From an evolutionary perspective, this has positive implications for the replicative strength of memes that possess this feature of 'violating ontological expectations'.

With regard to AI environments, there is reason to question whether AI algorithms qualify under this supernatural condition. Later in the chapter, Boyer draws a sharp distinction around what constitutes a violation and what constitutes an 'oddity' (2001, 81). A requirement for a violation is that all instances of the ontological template must be the same and possess the same qualities. For example, a person can have their ontological template violated because all humans have the same ontological framework. All humans are all made of the exact same matter (more or less), and are restricted by the same laws of physics. In essence we can say that humans are not multiply realisable (they only come in one form, or two if one wishes to separate the two sexes). Telephones, on the other hand, are only ever subject to oddities, because their constituent qualities are dynamic. If you open a telephone, it is likely to be made of different parts and perform different functions than the one you opened previously (unless they are the exact same make and model). To use the same terminology, telephones are multiply realisable. If you have a magical telephone, it is just that: a magical telephone, not an ordinary telephone with special extra 'magic' expectation violation.

This distinction places algorithms in an awkward category. As we have explored, the human mind has the tendency to anthropomorphise algorithms through the HADD. Does this process necessarily project the ontological framework on a person onto the algorithm (i.e. the algorithm can 'know', 'understand'. All features usually reserved for the 'person'

framework) in the mind of the projector, which could then be violated by an act of superhuman predictive capabilities, leading to an algorithm that is believed to possess some form of supernatural intelligence? Of course, this cannot be the case, I do not believe that there is anybody that truly believes that an algorithm is using some sort of supernatural power in order to arrive at its conclusion. Here are two solutions to this that debunks the idea of algorithms satisfying Boyer's conditions on supernatural concepts:

1. The easiest option is to deny the idea that an ontological framework is transferred through the HADD process, and therefore an algorithm does not qualify under the necessary conditions and is therefore not supernatural.
2. Alternatively we can choose still attribute agency to algorithms through the HADD, however we must admit that we cannot bound the ontological framework that is applied in the same way as we would if it was an individual human, because algorithms are multiply realisable and therefore we can consider an algorithm as being any number of humans, not just one. In this case, the performance of an algorithm is interpreted as an oddity, because it is compared with what any number of humans could accomplish, which is theoretically only bounded by the laws of nature (which, through a materialistic view, we can assume that an algorithm could not breach either). The conclusion being, again, that it is not supernatural.

So then, we can conclusively say that AI environments will not be able to count on the same level of recall for replicating ideas of their intelligence. However, it does still seem that these algorithms are (or have the capacity to be) extremely close to all-knowing and all-sensing. This does not specifically categorise them as being supernatural, but it is the closest comparison that can be found in religious concepts, which warrants a further exploration.

3.5.3 Psychologically primed: Supernatural as Opaque Complexity

Boyer's work signifies an abundance of accepted (and consistently recalled) supernatural concepts. He categorised them in order to dissect them and understand how they replicate

through human minds, but I wish to examine how they in-turn affect the same minds that replicate them.

For the purposes of examining common features across religions, it makes sense to categorise supernatural features together, especially, as we have just seen, in order to identify similarities and qualifying conditions. However, as Boyer briefly mentions later in the chapter, these supernatural concepts only hold up at one level of analysis (i.e. Q: What is God? A: He is an all-knowing and all-sensing man that is everywhere at once). At the next level of analysis, a concept that appeals to the supernatural is also an example of an opaque complexity that cannot be conclusively explained but must be accepted without further explanation in order for the enterprise to not crumble (i.e. Q: How can God be everywhere at once? A: We can't know, but if we take this as a reason not to believe then we are wrong). In essence, people invent these supernatural agents, and then degrade –or make exceptions in– their standards of explanation and acceptance so that the existence of these agents remains compatible with their existing ontological expectations. In this reversal of logic, the validity of the agent is never in question, only whether or not the person believes in them.

Jaron Lanier outlines a similar phenomenon that happens with technology, especially with regard to automated decision making, where humans consistently 'degrade themselves in order to make machines seem smart' (Lanier 2011, 32). By this, he means that humans will make allowances in the level of sophistication that they hold technology to, oftentimes by reducing their own intelligence, in order to preserve the clout of that technology. 'Did that search engine really know what you want, or are you playing along, lowering your standards to make it seem clever?' An example Lanier uses is bankers believing in supposedly intelligent algorithms that could calculate credit risks, even though they were leading them into a financial crash. This phenomenon is not just an outsourcing of decision-making to a seemingly reliable agent, but a voluntary degradation of intelligence in order to preserve current beliefs (e.g. the belief that the technology they are using is worthwhile). This is also documented by Jennifer Logg and her team at Harvard in their paper 'Algorithm Appreciation', in which they demonstrate that people are much more likely to choose the advice of a machine rather than a human (Logg, Minson, and Moore 2019). The subsequent result of this phenomenon is described by Lanier in his section

titled ‘The Turing Test Cuts Both Ways’ (2011, 31), and also by Frischmann and Selinger in their section titled ‘The Human Side of the Turing Line’ (2018, 179). Both outline the same concept, which is that we typically only ever focus on one side of the Turing Test, and fail to consider the sides being reversed. The side that we primarily focus on is whether a machine could ever convince a human into thinking it is actually a person (when compared directly against another human), however there is another side of the Turing Test, where the machine does not need to become more like a human if instead the human is engineered to act more like a machine.

While the comparison between the Turing Test and people that believe in the supernatural may not be an intuitively obvious comparison, here we see a cultivated (and rewarded!) mindset that willingly degrades its ontological standards in order to adhere to the common value system of the era. If that value system is an engineered environment where machines are thought to be intelligent (as would be the case in an AI environment, where intelligence is attributed to algorithms), then it becomes very plausible that people will degrade their ontological expectations of intelligence in order to make sure that machines really are perceived as intelligent.

AI environments may not be able to leverage the replicative advantage that religious concepts had where they employed claims to the supernatural, however they are able to leverage the by-product of this phenomenon, which is a human mind primed to accepting opaque complexity in systems, even when they violate their ontological expectations of that system, and even going so far as to reduce their standards in order to maintain the validity of whatever complexity is being concealed.

3.5.4 Psychologically primed: Personal Sacrifice (exchanges of value)

Self-abdication can also be seen in acts of religious sacrifice, whether material or internal. These actions manifest in forms such as voluntarily surrendering livestock to the gods, or making an internal sacrifice of mental and moral privacy in confession. With regard to material sacrifice, Boyer states “the justification of [sacrificial] performance is almost invariably the notion that misfortune can be kept away and prosperity or health or social order maintained if the participants and the gods enter into some mutually beneficial

exchange relation" (2001, 241). Here we can see a similar mentality to the previous example of people unconsciously sacrificing their intelligence in order to maintain social order, however the types of sacrifice outlined here are clearly explicit, with explicitly intended outcomes. In this context, a 'sacrifice' is more like a commercial exchange than a true sacrifice (i.e. an expenditure without return), however, as Boyer points out, this is 'almost invariably' the way that religious sacrifice is performed (i.e. as a 'mutually beneficial exchange relation'). Where vagueness may arise is whether maintaining the status quo could be interpreted as a return. If so, then most religious sacrifices would be seen as having no return, whereas (and under Boyer's interpretation), the maintenance of social order or avoidance of misfortune constitutes a return. In extreme cases, for instance Kierkegaard's argument that Abraham must be willing to lose it all for nothing, the definition of sacrifice used here would not stand. The definition of sacrifice used in this context is that of a mutually beneficial exchange, where the return is not fixed, largely opaque, and may constitute 'nothing' (i.e. no return) in the form of the maintenance of the status quo.

Similar mentalities exist in AI environments where the renunciation of personal privacy and personal data is encouraged under initiatives such as 'transhumanism', 'datafication' and slogans such as 'information wants to be free'. The 'mutually beneficial exchange' model for these initiatives is that a user should make all of their data (personal and behavioural) data free and accessible, and in return the user can avail of the short-term benefit of using 'free services' (which are able to run based on revenue streams from selling the data to third parties) and with the long-term promise of a digital utopia, often referred to as the 'Singularity', where –given enough personal data– each person's consciousness will be uploaded to the cloud in which people can escape all of the biological limitations that currently oppress them (such as hunger, animalistic desire, and mortality).

This strikes a remarkable similarity to religious concepts like the 'Rapture' (and other heavenly promises) prophesised by religions, that make promises of eternal salvation in exchange for the sacrifice of items of varying value in the present. These promises can be very convincing, especially considering that generally these sacrifices can make perfectly rational sense, even to a complete sceptic (for example, Pascal's Wager (Hájek 2018), a

pragmatic approach to religious belief as retaining the best cost-benefit ratio if accepted). At the extreme end, to a devout believer, even a sacrifice of their most valuable material possession cannot outweigh the value of eternal salvation. At the other end of the scale, to even the most dubious sceptic, most of the sacrifices encouraged can be held in seeming indifference. It is not a major inconvenience to have to tell a priest about minor sins even if heaven does not really exist. Likewise, it is not a major inconvenience that Facebook records minor details about what a person looks at and sells it to a third party, and all the better if eventually it contributes to an immortal digital consciousness. As far as cost-benefit analyses go, this trade-off can be rationally justified as consisting of the largest potential benefit on a cosmological scale, at a disproportionately insignificant relative cost.

These conditional promises for a digital utopia are still in the mainstream and acquiring proponents, even despite previous believers coming to the realisation of what is actually being built with their sacrificial data (essentially giant marketing and behavioural modification machines that make huge profits for their owners at the sole cost to the user) and voicing their disapproval for the current business model; such as the subtitle of Founder Fund's manifesto *What Happened To The Future*, that states 'We wanted flying cars, instead we got 140 characters' ("What Happened to the Future?" n.d.), referring to the character limit of tweets being extended and Twitter broadcasting the change as if it constituted technological progress.

3.6 Final evaluation of AI environment concepts.

To summarise this section, we can see how AI environments stand to capitalise on existing biological features that were also capitalised on by religion. It is reasonable to speculate that these features were further intensified by generations of recursive reinforcement by common religious concepts, by intensification in either the gene-pool (as speculated in the case of shamanic healing) or the memeplex (for example, in the case of ritualised behaviour).

From then moving on and exploring the cultural evolution of religion, it became evident how these memes further capitalised on 'misfiring cognitive modules'. This process leveraged the human mind being 'psychologically primed' to replicate certain concepts, provided they appealed to the correct mental modules.

At the beginning of this section, four necessary conditions were introduced which would help evaluate whether a collection of features resemble a religious concept. To recap, these conditions were:

- Making recall and communication easy
- Triggering emotional programs
- Connecting to our social mind
- Becoming plausible and direct behaviour

To conclude, I will evaluate each of these points with reference to AI environments.

3.6.1 Making recall and communication easy:

Religious concepts cultivated and maintained a social contract between members through the use of ritualistic behaviour in announcing changes to the contract, they encouraged recall through ontological violations in their features, and through rigid ritualistic practices ensured that they maintained consistent communication across members and with their deities.

AI environments primary platform for data-gathering and content-dissemination is through social media platforms. These platforms encourage communication across members, who are linked together through mutual defining features (such as common hashtags, recommendations, identities, subcultures). These groups are typically automatically generated based on behavioural modelling and are disseminated in the form of recommendations and automatically curated news feeds. Changes to the social contract are encouraged, and they aid in group harmony, as expectations of each member can be continually managed in the minds of other members. This also benefits the automated decision-making systems that rely on this data to accurately model the user in order to make the most accurate predictions about them and their values.

Rigid ritualistic practices are also engrained in the user interface and limited available choices for interaction on these platforms. To the uninitiated, ritualistically scrolling through Twitter on a morning commute must look remarkably inexplicable, comparable to observing someone participating in the thumbing of prayer beads. This rigidity holds benefits for the machine-learning algorithms that drive AI environments, as simple, consistent, and vast behaviour datasets are generated from this heavily constrained level of participation.

Unfortunately for the algorithms, it seems doubtful that they qualify as ontological violations, and thus cannot rely on the levels of anecdotal recall that supernatural agents had. However, it seems reasonable to believe that these algorithms will still stand out as memorable to users due to the HADD and a hypersensitivity to personal-level interactions.

3.6.2 Triggering emotional programs:

Religions trigger emotional programs through ritualistic behaviour and sacrifice, and personal level interaction. As has been explored, ritualistic behaviour is typically accompanied by a compulsion to perform it, especially with regard to sacrifice in so far that ‘misfortune can be kept away and prosperity or health or social order maintained’ (Boyer 2001, 241) and the sense that the believer is emotionally invested in the endeavour is often more important to the success of the ceremony that the actions themselves. Religions also evoke emotions through personal level interaction, where the subject comes

to believe that they are the centre of the event and that a supernatural agent is communing with them and them alone.

This level of personalisation and emotional investment is paramount to the success of AI environments, as personalised information is the commercial imperative of the platforms that are used. A sense of personal understanding and interaction is the main allure of these algorithms, and why they are attributed with so much power. These algorithms are given power by users to make decisions on their behalf because of the sense of personal understanding that they (seemingly) exhibit. Some of this is valid, but some of it is wishful thinking on behalf of the user and just an illusion created by the HADD.

In addition, emotive behaviour can be observed with the obsessive mentality that users have with their connected devices and ritualised behaviour around updating profiles and maintaining an active presence in these algorithms. There is an implication that the algorithms resemble the user, and if these algorithms are not maintained with the current values of the user, then they become stale and misrepresent the person. This idea of 'pollution' is repulsive to the user (who does not wish to be wrongly represented and misunderstood) and to the platform (who do not want to model the wrong behaviour) and so measures are taken by both parties to ritualistically cleanse and update the datasets, and also to create environments where that behaviour is continually encouraged ('Hey, you haven't posted in a while').

Further to this, it is evident that these behaviours create very intense emotional links between a person and their actual physical devices, as explored with regards to compulsion, separation anxiety, and nomophobia (Rodríguez-García, Moreno-Guerrero, and López Belmonte 2020; Kuss and Griffiths 2017).

3.6.3 Connecting to our social mind:

At the earliest conception of religions, at the level of archetypes, it is evident that religions connected people. In this sense, Yuval Noah Harari takes a functionalist interpretation of early religion as a means of mass social cooperation outside of traditional narrow tribes. The subsequent evolutions of religious concepts continued this tradition of creating common systems of values, as a means of understanding who constituted 'us' and who

constituted ‘them’. This had benefits internally, with regards to social harmony, but also provided framework for the de-humanisation of ‘them’ - also ‘other’ ref Husserl & co- groups where they could be seen as holding contradicting values, which further affirmed internal relations (i.e. “We are definitely right, even despite minor differences, because they are *so* wrong”) and gave moral licence for wars, conquests, and slavery, all in the name of ‘good’.

Similar methods of maintaining social harmony can be seen in ritual behaviour, where social contracts are continually updated through ceremonies, which maintains understanding of expected behaviour from each member.

Comparable examples of insular feedback loops can be seen in AI environments, where users with similar profiles are grouped together, shown similar content and generally reinforce one-another’s opinions. A difference, however, is where conflicting value systems clash in these environments, in the absence of intermediary figures that religions have (such as priests and guidelines on how to deal with conflict, such as ‘turn the other cheek’) then public discourse declines where both sides are convinced of their value-system, due to such a large group of similarly profiled people. Some technologists approve of this phenomenon by making appeals to the ‘hivemind’ of the internet. In either case, it is a contrasting point that shows how religions put in place means of resolving (or dissolving) group conflicts by having prescribed value systems and mediators. Further generations of AI environments will likely need to develop some kind of overarching value-system that can be called upon to handle disputes (this idea of an encoded value-system guiding the decisions of AI environments is explored further in Section 3). This is not to say that these environments do not ‘connect to the social mind’, conflict is still a social act. In fact, it has been proven in a controversial study (ethically controversial, not controversial in its findings) that emotions are contagious through social media feeds, and that the more negative posts that a user is subjected to, the more negative their posts become, and likewise vice versa (Kramer, Guillory, and Hancock 2014). This, combined with the fact that people interact more when they are exhibiting negative emotions (Fan, Xu, and Zhao 2016) leads to an incentive for these platforms to manipulate feeds to induce more emotional conflict in order to encourage more interaction, leading to more behavioural data and more revenue in the attention economy.

3.6.4 Becoming plausible and direct behaviour:

Ritualised behaviour plays a significant role in the propagation of religious concepts. Religious concepts are able to maintain perfect fidelity across transmissions through strict rules around replication (i.e. just muttering a few sentences is not a prayer, it must be verbatim) and at times leveraging opaque complexity so that even intentional mutations cannot occur (i.e. if a prayer must be repeated in its original language, even if that language is not commonly understood any longer). With ritualistic behaviour being associated with anxiety states and obsessive compulsions, religious concepts were able to become mental crutches that people relied on to maintain a sense of social order and normality. Rituals gradually became normalised with each performance, eventually becoming a normal behaviour to the initiated, despite seemingly obvious displays of goal-demotion.

In much the same way, we see how contracting and surveillance creep can lead to a gradual habituation to ritualised behaviour in AI environments, and can lead to passive information seeking behaviour along with an implicit and unexamined trust in intermediary parties and authority figures.

Behaviour in religions is not just limited to operating within the context of the religion. We see intertwined processes emerge as they integrate with society, for example, getting married in the eyes of god also has very real consequences for state tax, namesakes, and other societal obligations. In much the same way, we are beginning to see user profiles becoming legitimate representations of people in society, for example LinkedIn profiles being used as resumés in job interviews. This brings along with it the societal (and potentially legal) obligation to keep these datasets continually up to date.

3.7 Conclusion.

This section explored how archetypal value-systems moved from being decentralised value-systems to being curated centralised collections of myths, and then onto being religious concepts. This transformation was initiated with the introduction of intermediary

parties that gave the concepts increased replicative strength over their existing strength in social cohesion and internal resonance. The exploration began by documenting how certain biological traits impacted the evolutionary fitness of religious concepts, and what necessary features these concepts all share. This process of evolutionary strengthening was also heavily influenced and amplified by ritualised behaviour and the human proclivity to participate in it.

Later stages of the section examined how religious concepts came to depend less on human biology, and more so on human mentality, and at that point departed from their extended-phenotype classification and became easier to analyse as ideas in the memeplex. From this point onwards, it became easier to document how religious ideas capitalised on the human mind through leveraging cognitive mis-fires. Through this exploration, a detailed list of mental traits emerged that appeared to account for aspects of the human proclivity toward religious concepts.

The final evaluation of this section consisted of a comparison of features that exist in AI environments. These features were categorised under the list of mental traits that religious concepts has curated and leveraged. This evaluation determined that it is likely that the human mind has the capacity and proclivity to interact with AI environments in similar ways that they did with religions.

4 Speculation About The Future.

4.1 Reduced Dimensions of Experience

The primary concern in relation to the future of human interactions with AI environments is that dimensions of human experience will be reduced. The first section of Lanier's manifesto, *You Are Not a Gadget*, addresses some of these concerns, regarding internet culture and how gradual habituation to a 'common standard' (15) will result in humans limiting their own experience to what can be represented in a database. He uses the 'lock-in' of MIDI as an analogous example of musical experience becoming 'over defined, and restricted in practice to what can be represented in a computer' (10), which –given a 'common standard'– could happen to different facets of human experience.

Lanier's concern deals primarily with internet culture and technological encoding in general. The sentiment of his manifesto is that insufficient standards will arise, that do not accurately represent human experience, and over time, humans will gradually degrade their expectations to these standards. This gradual habituation could lead to humans not exploring certain areas of creative experience. My concern regarding AI environments is similar, however, I am concerned that AI environments will lead to the outsourcing of the exploration altogether. For instance, Lanier worries that humans will not explore certain spectrums of sound because their expectations of sound are limited to what is available within the MIDI standard, whereas I am concerned that humans will not even explore the MIDI spectrum, but rather only experience the sounds that are recommended to them by an intelligent system.

My expectation is similar to a concept introduced by Shoshana Zuboff, which she calls 'instrumentarianism' (8). Instrumentarianism refers to the manipulative method of behavioural modification that some early AI environments are used for, especially in advertising, where it has become easier to change what a person values instead of trying to predict what they do (or will) value and to cater to it.

In religious terms, it is easier to use the 'voice of God' to promote an agenda than to adapt that agenda to integrate with existing values.

The difference between Larnier's MIDI example and my expectations can be reduced to a distinction between overdetermined and constrictive environments. Frischmann and

Selinger introduce and explore this distinction in a section exploring 'Engineered Determinism and Free Will' (228). They outline that 'overdetermined environments … eliminate the practical freedom to exercise will by constraining the range of actions or opportunities presented', but 'constrictive environments … engineer the will by determining beliefs, preference, tastes, or values'. They suggest that, at extremes, overdetermined environments can lead to slaves, where the environment dramatically reduces the scope of opportunities (like MIDI did to the entire musical spectrum, where musicians became 'slaves' to the MIDI standard). Whereas it is suggested that constrictive environments lead to 'simple machines' which is, in essence, the 'degeneration of autonomy into simple, stimulus-response behaviour by humans' (227).

To keep the theme of musical examples in this section, let us imagine a constrictive AI environment that recommends songs to users based on the preferences of similar users. The objective of the environment is to recommend songs to the user that they are likely to listen to. In this environment the user's actual preference of each song (as well as their overall musical preferences) will go largely unrepresented, other than through the narrow Key Performance Indicator of accumulating 'listens'. While it is likely an effective strategy for recommending songs that the user is likely to listen to, in this scenario humans become functionally equivalent to song-listens-count machines that convert music consumption into revenue. With this flawed reduction of preferences to song listens, it is easy to mistake the map for the terrain and skip the user preferences altogether. A recent example of this involved Justin Bieber asking his fans to listen to his new music while asleep, or to just listen to his new song on repeat even with the speakers turned off, in order to get his newest single to number one on the charts. He even asked them to use a VPN to make it look like they were listening in the United States, so that the streams would be counted against the US Billboard charts (Tenbarge n.d.). This is an example of the proxy for good music (number of listens) being confused as what constitutes good music (being a song that people *want* to listen to), which brings about the impression that one can somehow engineer subjectively good music by increasing the number of times it is heard. The consequence of this process is that it reduces human involvement to that of a stimulus-response machine whose independent subjective experience is largely irrelevant.

To summarise the current standing, Lanier is mostly concerned with humans becoming slaves: they retain the capacity to be free, but their lives and the opportunities afforded to

them are overdetermined to the extent that they never exercise this ability and willingly accept their narrow range of experience as the full breadth available. Living in a 'locked in' society could have significant implications for moral responsibility, which is a necessary by-product of free-will that plays an important role in self-regulation and expectations of personal responsibility within society. As illustrated in the literature review in section 1 of this thesis, Harry G. Frankfurt outlines that the 'principle of alternate possibilities … [which] states that a person is morally responsible for what he has done only if he could have done otherwise' (1969, 829) is not sufficient for moral responsibility and free choice. Beebee formulates an alternative to this principle, in the form of her 'Principle of Unforced Action' (PUA) that necessitates a condition that it is also not the case that the person acted in a particular way *only* because they could have not acted otherwise (Beebee 2013, 141). Essentially, an agent is only morally responsible for an action if it is what they chose to do, regardless of whether or not they *could* have done otherwise. In a locked-in society, the range of human experience could be voluntarily reduced or over-determined to the extent that citizens no longer retain the ability to choose otherwise because they are not aware that other options exist, which ultimately undermines their level of responsibility. This shares a political comparison with Étienne de La Boétie's manifesto on 'voluntary servitude', in the introduction to which Murray N. Rothbard summarises that 'in the beginning men submit under constraint … but those who come after them obey without regret and perform willingly what their predecessors had done because they [i.e. their predecessors] had to. This is why men born under the yoke and then nourished and reared in slavery are content' (1975, 21). In this scenario, an overdetermined environment is essentially 'locked in', and expectations of experience are reduce to fit it, however it is not necessarily the case that the 'slaves' lose the capacity to conceive of other states of experience but rather they do not have access to them, or perhaps they just do not desire them. In these scenarios it becomes increasingly difficult to say that these people are exercising free-will. An example of this can be taken from Frederick Douglass' *Narrative of the Life*, in which he discusses slaves arguing over who has a relatively kinder master (1995, 12). One can presume that a slave arguing in favour of their kind master is not incapable of conceiving of a value-system where they are not enslaved, but rather they have reduced their expectations of experience to that of a slave and framed their value-system within that limited scope.

4.2 Developing an Inability to Self-Determine Values.

My concern is that the slave discussed at the end of the last section could *become incapable* of conceiving of a value-system in which they are not enslaved, provided enough generational iterations of outsourced cognition, that is to say: permanently losing the ability to explore and evaluate alternative scenarios which becomes an inability to self-determine values, which subsequently undermines their free-will. A good contemporary example of this is losing the ability to way-find. This does not necessarily have to constitute finding one's way through an unchartered rainforest, but something as simple as finding an alternative route across town. With the increased use of applications like Google Maps, that can effectively plan routes exceptionally better than any human (or at least, much faster), it is understandable that people choose to use them as reliable outsourcing tools for wayfinding. However, when outsourcing this task, one is sacrificing the opportunity to develop the skill of exploring alternative scenarios, i.e. evaluate different potential scenarios for 'truth'. In this instance, the fastest route to the destination is the 'truth'. By outsourcing the task, one sacrifices developing all of the mental processes that are involved in 'truth'-finding and reducing them to the simple process of –at best– evaluating the suggested route times and selecting the one with the fastest travel-time, which is as literally trivial as selecting the lowest number from a short list. This neglect in developing this skill, followed to its conclusion, can only lead to the permanent loss of the ability to explore alternative routes (other than wandering around aimlessly), which can be understood as the inability to self-determine the 'truth'. While this is a limited case referring to the ability to way-find, it acts as a contained example for the initial rationality and process behind outsourcing the ability to determine values (i.e. the necessary skill to way-find in a conceptual space) which, if applied to something like morality, would have significant implications for each individual in their development of the skill of self-determining values. This idea of under-developing a mental function is addressed, in part, in William Poundstone's book *Head in the Cloud*. The author uses numerous studies to illustrate 'how much we are coming to depend on source memory' (2017, 28) ('Source memory is the recall of when or where a fact was learned. It is often fallible and has been implicated in false memories' (28)). As an example, in one Harvard study, subjects were provided with a list of trivia facts, and told that they would be stored in a certain folder.

These folders were given generic names, such as ‘data’, ‘info’, ‘facts’. The trivia-facts, instead, were ‘quirky and memorable’ (28). The results of the study found that ‘volunteers were more likely to remember which folder stored the trivia facts than the facts themselves’ (28). These types of studies provide empirical evidence that interacting with these systems is influencing the way our brains store and access information. While it is not that this is necessarily making us stupid (as is frequently argued to be the case (Carr 2008)), it is also not something that one should be considered a luddite for being concerned by because it *is* having some kind of effect on our mental development. As the author summarises: ‘The great risk isn’t that the Internet is making us less informed or even misinformed. It’s that it may be making us *meta-ignorant* – less cognizant of what we don’t know’ (26). We might not develop cognitive skills because we presume that relevant information/ability will be there if/when we need it.

Let us consider philosopher Robert Howell’s idea of a fictional application called ‘Google Morals’ (2014) in order to explore this concept of ‘under-developing a cognitive ability’ in what is generally considered to be a more intimately individual process than trivia fact recall or way-finding. ‘Google Morals’ acts in the same way as Google Maps in the sense that it helps a person navigate terrain from a starting point (current reality) to a desired end (morally virtuous destination), except it navigates a moral landscape instead of streets. The assumption driving this thought experiment is a claim that morality is computationally understandable, which Howell does not actually believe to be true. While it would be a very interesting (and tempting) tangent to make, I will not discuss whether morality is computationally understandable, because for the purpose of this research, all that is required is that the user wholly and undoubtably believes that Google Morals is an infallible source of moral wisdom. There is reason to believe that this is plausible, for two reasons:

1. Religious texts claim to be infallible sources of moral wisdom, and that was undoubtably believed (and still believed in places) by their users. The previous sections have identified reasons and provided an evolutionary account for this, regarding both sides of the partnership (i.e. partly due to system design, partly due to user-mentality).
2. Techno-utopianism: there is a common-held belief amongst techno-utopians (i.e. those that believe that sufficient technological progress will bring about a utopia)

> that what is being created by the hyper-connected population of the world is a type of hive-mind (https://kurzweilai.net n.d.). This hive-mind, they claim, is a type of self-organising system of collective human values, which lives as a distributed organism across the internet and connected devices. This hive-mind would represent a universal encoding of morality.
>
> Recent studies demonstrate how 'we how come to "own" the Internet as collective memory' (Poundstone 2017, 33). Across two studies, subjects were asked to answer trivia questions and were permitted to look up the answers online. After the quiz they were asked to rate their 'memory, knowledge, and intelligence' (32). Participants that scored high in the quiz rated themselves as being smart, however, 'the eye-opener was that the ratings were higher for those who had just looked up everything' (32). This implies that these participants saw the knowledge contained in the 'hive-mind' as just a collective memory, to which they had a rightful stake.
>
> This idea of a self-regulating and self-organising harmonious system is also found in key tenets of numerous value-systems, such as Friedrich Hayek's 'spontaneous order' of free-market capitalism (Schmidtz 2019), Buddhism's belief in karma (Lopez 2004, 24), or Taoism's 'middle way' (Watts and Huang 1975). The human desire for an unattended natural law that maintains a cosmic balance (whether economically, morally, or spiritually) can be psychologically understood through areas of behavioural economics such as loss aversion, altruism and punishment (Clavien and Klein 2010). These theories explain the innate human need for fairness (or at least the belief in a universal equaliser), and we can speculate that humans will therefore look for one in any theory of economy, religion, or algorithm.

In this sense, we can assume that techno-utopians will promote the hive-mind's understanding of universal morality, and that the majority of users will come to believe it, and in some cases, even seek to believe it in order to reassure them of a universal fairness.

Frischmann and Selinger address the idea of Google Morals in their book, however their concern extends as far as people being afraid to not use Google Morals due to its infallibility and the demand on moral cognition that could be easily outsourced (230-231).

However, I do not think fear sufficiently captures the interplay that is likely to happen in this scenario. A self-reflective emotion like fear stops slightly short of fully articulating the way that users are likely to interact with this type of system. I do not believe that users will maintain this sense of self-awareness of being fallible creatures and will therefore not feel fear at the thought of making their own decisions, but rather that they will not be aware that there is any other way of making a decision. To be able to identify self-doubt requires a self-awareness that is not encouraged in these systems. Users would not turn to the system because they are consciously aware that it is better at making decisions than they are, they turn to the system because *that is how they make decisions*. The system becomes the development of the decision-making process. I propose that eventually the user would not go through a complex emotion of being afraid to make their own decisions, but rather not reflect on whether there was a different way of navigating a moral landscape at all (i.e. there is no source of absolute truth other than the algorithm). The perception would be that Google Morals is a comprehensive map of the value-system of the world, in exactly the same way that religious texts claim to be. Given this perception, to question a single assumption made by Google Morals would be to, by definition, be wrong (in the ultimate sense of 'Truth').

It is reasonable to expect that AI environments like these will gradually instil a passivity toward information evaluation, as is explored in the Google Morals example. This, over time, could potentially cultivate a mentality that becomes incapable of self-determining values, as the tendency to outsource moral choices becomes easier and the internalisation of these choices becomes (by contrast) increasingly difficult.

4.3 Becoming Stimulus-Response Machines & The Loss of Individuality.

This concept of technology-induced passivity can often be misunderstood. Popular culture often likes to sensationalise the idea that a continual integration with technology –and especially an over-reliance on technology– will ultimately turn humans into mindless consuming robots that become overweight, sit in a hover chair consuming fast food and mass media, being kept alive only as a means of propagating a machine (such as Pixar's *Wall-E*, or even E.M. Forster's *The Machine Stops*, albeit lacking the Americanisms). I do not believe that this will be the case. If anything, I believe that humans will become ever

more active, in much the same way that many increasingly devout religious believers become increasingly active in attending ceremonies, enacting rituals, and propagating the message of God to non-believers (sometimes even to the extent of mobilising entire armies for a crusade in the name of their chosen deity). I believe that hyper-connected humans that outsource their value systems will not become passive in a sense of aimlessness and meaningless action, but passive in the sense they will not evaluate information sufficiently enough to make decisions that they can consider their own, which ironically, could result in their actions being even more detrimental than if they were just passive in the sense of meaninglessness.

If humans become passive in their evaluation of information sources then there is reason to believe that the dimensions of human experience will all converge toward an average. This constitutes a technical issue with AI environments, where recommended values cannot sufficiently cater to unique individuals, by definition. This seems paradoxical, because the premise behind these AI environments is that content is personalised to the unique user. This is not the case. If we remind ourselves that values are suggested based on a lookalike profiling model, then it becomes clear that values are not personalised to fit the person, but rather that the person is positioned to fit the value. Take for instance a political campaign where the candidate has 140 different proposed policies, 5 of which are shown to the voter depending on their profile. The 'value' in this instance is the decision whether to vote for the candidate. In this scenario the suggested value is not personalised to the voter (i.e. the value is not dynamic, the candidate does not change based on who the voter is) but rather the voter is positioned to fit the value through a process of averaging (i.e. the average voter of this demographic will adopt this value under these conditions). This may be an example that is too actively deceptive, however. Instead, consider simple content platforms such as Twitter that curate news feeds for the users based on behavioural data. It is widely accepted that these sites create feedback loops, where similar content is shown to similar minded users (e.g. the average millennial user wants to see this piece of content) (Garimella et al. 2018; Bastos, Mercea, and Baronchelli 2018; Bodó et al. 2019). Through this process of averaging behaviour and curating content to fit the profile of the user, this group is encouraged to converge closer to the average of their group. Over iterations of these suggestions and behaviours, outlying content will be gradually reduced as the relative score of this content will always be insignificant in comparison to the level of interaction

that the most-generally-engaging content receives. As people using these sources of information continue to consume suggested content instead of actively searching for content (which is why people use recommendation systems in the first place), then they will tend closer and closer towards a monoculture where everyone in a certain profile is shown the exact same content, and recommended to accept the same types of values. This encouragement towards average, combined with a readiness to trust these systems as reliable sources, could condition humans towards becoming stimulus-response machines in engineered environments.

Jaron Lanier compares this move toward globalisation and monoculture to 'missionary reductionism' (2011, 48). In this section, he explores how 'strangeness is being leached away by the mush-making process' (48) (by 'mush-making process' Lanier refers to the 'digital flattening' (45) of content, through processes such as authorship removal, aggregation sites, and crowd dynamics (i.e. feedback loops, filter bubbles) previously discussed). Lanier uses an example of how 'elements of indigenous cultures were preserved but de-alienated by missionaries'. In essence, he describes how new content is created or discovered and in the process of digitisation and distribution the outlying bits are trimmed and discarded, with the surviving content consisting only of the sections that suitably fit into the existing general conception of what that content should be. He emphasises that the bits that get cut off are typically the most 'precious' bits, as they are the pieces that create diversity, or in his words 'portals to strange philosophies'. While Lanier writes with specific reference to individual representation through technology (e.g. all social media profiles having the same layout with 'multiple-choice identities', and Aztec music being 'trimmed to make the music fit into the European idea of church song' (48)), it can be extrapolated to suggested content, as the only content created is created under the same template. Lanier draws a comparison between religions with their desire for having just 'one book' with no authors (i.e. no individual points of view, just one universally averaged value-system) and the similar desire of certain Silicon Valley figures to eradicate authorship in favour of creating a 'universal computation cloud' (46). Lanier claims that '[a]uthorship–the very idea of the individual point of view–is not a priority of the new ideology' (47) due to the fact that content platforms encourage aggregated content and condensed versions where 'considered whole expressions or arguments' are largely irrelevant because of the interaction-based economic model that powers them. This has an

immediate comparison with religious texts, in which individual authorship is entirely bypassed and replaced with concepts of deities communicating *through* the authors.

Yuval Noah Harari echoes similar concerns for individuality in his book *Nova Deus*. He predicts that in a near future, powered by technology and intelligent systems, human individuality will not be forcefully suppressed by a genocidal dictator, as it was in the twentieth century, but rather that 'human individuality is now facing an even bigger threat from the opposite direction. In the twenty-first century the individual is more likely to disintegrate gently from within than to be brutally crushed from without' (2016, 402). This is a reality that we already see happening in the era of fake news, information overload, and mass confusion (Poundstone 2017; Wylie 2019), where staying informed constitutes a choice between being misled (accepting information received at face value) or overwhelmed (attempting to continuously critically evaluate a bottomless newsfeed full of incomplete arguments). The individual is soon reduced to a stimulus-response machine or alternatively is crushed into mental paralysis in the form of a constant state of doubt. In either case, they can no longer claim to be an individual in the sense of freedom as a self-determining, morally-responsible agent.

4.4 An Environment for Individuality.

If it becomes the case that users in AI environments become passive and incapable of determining their own values, then we must ask, are they free? If turning to a system is the new way of making a decision, perhaps it has just become the first step in the decision-making process, rather than replacing the whole process altogether? Is there still a chance that humans could self-determine their values, even though their decisions are not made inside their own minds? We will answer these questions through an exploration of the theory of 'extended mind' and by outlining a potential future involving AI environments becoming the tool used by humans for fulfilling Kant's categorical imperative through self-determining values and retaining moral responsibility.

At this point it is important to outline what constitutes a desirable future. What necessary conditions must be met in order to say that one future is more desirable than the other? I propose that a Kantian future in which each person is a self-determining, morally

responsible, free agent is one worth pursuing. The necessary criterion to achieve this future is that humans must retain individuality (i.e. the sense that my idea is *my* idea, and not from without) as a necessary component for self-determining values. In order for this to be the case, a value that is given and adopted without examination, would be in direct conflict with this necessary rule. This potentially places AI environments in a freedom-denying category, as it appears that their only function to the user is to outsource certain cognitive processes. But perhaps this may not be the case. Earlier in this section, I questioned whether AI environments constitute a replacement of the decision-making process, or rather just become an additional step *in* the decision-making process. In order to establish whether AI environments should be placed in a freedom-denying category we must evaluate whether a decision made by a third-party agent on behalf of an individual can be considered as equally valid as an internal decision made by the individual.

There are a couple of key phrases to unpack here.

- **'a decision made by a third party *on behalf* of an individual'**: this evaluation will assume that the decision being made is a tailored decision, based on the agent's conception of the individual's preferences. This constitutes giving the benefit of the doubt to the agent in so far as we will assume that they will not opt to shape the individual to the value, as previously explored in the case of political campaigns.
- **'considered as *equally valid* as an internal decision made *by* the individual'**: in this instance, equally valid is not to say that an external decision is equally valid to an internal decision because it is also a decision, but rather that it must be the *same* decision that the individual would have made, given the same cognitive ability. We can suspect that an AI environment would have a far superior cognitive ability, so this is a difficult comparison to evaluate for 'sameness'. For instance, perhaps the individual does not understand a decision made on their behalf, but it *is* the decision they would have made *if* they had the same cognitive ability, they would just never know. Lacking the cognitive ability to make complex decisions is a prominent motivation for using AI environments, so it can be assumed that this will occur frequently. In these instances, it would be best to look instead at what is *guiding* the decision-making process, instead of what is *driving* it. Instead, we can consider sameness to be characterised by the morality guiding the decision-making

process, in instances of cognitive inequality. This is where computer-realisable morality becomes necessary. If we can confidently say that an AI environment has correctly modelled an individual's 'moral landscape' (a term taken from Sam Harris' book *The Moral Landscape*, which proposes a neuro-scientific approach to encoding a universal morality (2012), which may be a viable route), then we can assume that it would make decisions that are the same as those of the individual, if they had the same cognitive ability.

4.5 Can an AI Environment Qualify as an Extension to the Mind of the Individual?

With these criteria in mind, we can begin to explore whether a third-party agent can be the genuine source of a decision made on behalf of an individual. Philosophers Andy Clark and David Chalmers explore a similar concept in their well-known 1998 paper, 'The Extended Mind'. In this paper, the authors propose that the mind is not restricted to the individual's brain and that it can exist as a 'coupled system' with the environment it is in, through what they call an 'active externalism' (7) ('active' because of the 'active role the environment in driving cognitive processes' (7)). An example they use is of an individual playing Tetris evaluating where a shape will fit by rotating the shape in the computer game using a button instead of performing the task mentally (because it takes longer to perform mentally, as it so happens). In this instance, the cognitive task of rotating the object is outsourced to the machine, however the authors would suggest that the phrasing used there was superfluous for explaining the action. They say that these external cognitive processes 'demand … *epistemic credit*… [because] were it done in the head, we would have no hesitation in recognizing as part of the cognitive process' (8).

This example strikes an immediate contrast when compared with the decision-making that would take place in an AI environment. It is not the case that an individual would be using the AI environment in the same sense that someone would use a machine to rotate an object. If we were to say that, yes, cognitive tasks can be outsourced to a machine, then must we also say that we need some kind of internal governing body in order to say that it is actually a 'coupled system'? Say the Tetris player actually pressed a button that not only

rotated the shape, but that also decided the optimal place for the shape to go and moved it there, and following that the individual can only watch as the shape sits into place, at which point the next shape appears and the individual presses the button. Would we say that the individual was playing Tetris? Probably not. We could generously say that *they*, as in, the coupled system of the individual and machine, were playing Tetris. However, I am hesitant to say that this would be an *extended* mind, I think it would be more appropriate to call it a *distributed* mind (a distribution of which the individual's mind is doing very little). This example does not seem to allow much scope for a sense of self. At best, it seems like a joint venture between the two systems. At this point, we cannot really conclude whether an extended mind would facilitate humans in self-determining their values.

Perhaps we should consider a case where the mind outsources a *decision*-based task to an external agent (well, as we will see, an *internal*-external agent): The corneal (i.e. blink) reflex is a localised trigger and event (Esteban 1999), in the sense that the trigger for an reflexive blink does not come from the brain, but rather from a localised circuit. If we were to adopt Daniel Dennett's intentional stance introduced in the literature review, we could say that this localised circuit 'decides' whether to blink when it 'believes' that the eyeball is in danger of being hit. The reason for this is because it would take too long for the eye to identify a projectile, warn the brain, and then transmit the response to back down to the eye in time to block the eyeball from an incoming projectile. Of course, a non-intelligent system like this means there are frequent misfires when the system miscategorises something as a threat, but it is a minimal cost to pay for an efficient reflex when needed. The question here is whether this counts as an external decision-making agent? I think we can say that it is, in the sense that it makes a decision for the individual that is the same as they would make, given the circumstances. However, I would hesitate to say they are a 'coupled system'. The individual has no control over the decision or the subsequent action. There is no ability to veto by the mind of the individual (as you will know well if you have ever had an eye exam where they blow puffs of air in your eye, it is desperately hard to will your eye to stay open). In this instance, again, I would choose to categorise this as a distributed mind. We can consider this as a joint venture between the mind and an unconscious reflex that co-operate toward a desirable end, but not an extended mind that constitutes self-hood when making a decision. We can say that it is definitely a decision made by a third-party on behalf of the agent, but we cannot necessarily say it is equally

valid as a mind-made decision, because it is consistently wrong, where the mind would not be (for example, noises above 40-60dB can trigger the corneal reflex (Garde and Cowey 2000). No conscious mind would decide to blink in response to a loud noise).

So far we have dealt with examples of cognitive processing, where individuals rely on the environment to assist in bearing the weight of some cognitive load. The authors move on to discuss areas more associated with the mind, such as beliefs. They use an example of Inga and Otto, where Inga stores her beliefs in her head, and Otto (who suffers from Alzheimer's disease) stores his beliefs in his notebook. The thought experiment reaches a logical conclusion that Otto's notebook should be considered as equally valid of a repository of beliefs as Inga's internal system. This is an example that could apply to AI environments, where they could be repositories of trusted beliefs generated by an external system. While this is not the case in the example of Inga and Otto (Otto's notebook does not generate the beliefs inside it, but if it *could* then we could assume Otto would come to trust it, if it had a proven track-record). As the authors state: 'What is central is a high degree of trust, reliance, and accessibility' (17).

Funnily enough the authors doubt that the internet could ever act as an external belief repository: 'The Internet is likely to fail on multiple counts, unless I am unusually computer-reliant, facile with the technology, and trusting' (17). Ironically, I think this research has shown all three of these stand a high chance of actually being true (although it should be recognised that the paper was written in 1998, at a stage where the internet did resemble a repository of facts like Otto's notebook, and at a time when the idea of being able to carry supercomputers in our pockets–or even attached directly to our brains–would have seemed slightly outlandish). What is interesting about this, however, is the recurring theme of a central governor or 'trust'er that evaluates the work done by the external portion of the extended mind. This theme continues toward the end of the paper where they briefly explore possible 'repositor[ies] of beliefs' (18) such as 'the waiter at my favourite restaurant … act[ing] as a repository of my beliefs about my favourite meals' (17-18). They do not flesh out this use-case, which is unfortunate considering that it is actually the most applicable in this instance. Earlier I pointed out that in areas of cognitive inequality, that understanding the *guiding* influence might be more critical than the *driving* influence behind the decision. Using the existing characters of the waiter and the individual could

provide an appropriate example:

The individual is a busy surgeon who is largely clueless about food other than being able to identify what tastes nice and what does not. The waiter on the other hand is a master of taste, with numerous degrees in culinary arts, with a deep and varied understanding of complementary food tastes. In fact, the waiter has a special gift for understanding correlations between observing taste and deducing the aspects of the dish that guided that reaction. In this scenario, the individual goes to the same restaurant every day, and orders from this same waiter. For the first few weeks the individual orders her own food from the set menu and is either disgusted or thrilled by the dish that she receives. Of course, the waiter is conscious of her reaction, because he is either told that the food was disgusting or is told to give compliments to the chef. Additionally, he discovers that he gets a bigger tip if the individual enjoys her meal. Now that there is a potential reward in store for the waiter if the individual enjoys her meal, he begins analysing what tastes the individual has, and starts making suggestions based on these observations. More or less instantly the ratio of good to bad meals goes from about 50% up to the high 90%s. The individual is delighted, she gets to save even more time by not even having to read the menu anymore, and the waiter is receiving larger tips 40% more frequently. There are still some occasions where the suggestion is wrong and the individual does not enjoy the meal, but that gets fed back into the waiter's observations and only improves his accuracy the next time. And so, we must evaluate, is the waiter an extension of the individual's mind? Is the knowledge that the waiter possesses somehow an extension of the individual's knowledge? If the waiter just remembered a list of the individual's favourite dishes, then perhaps we could consider them an extension, because that list could be written down and taken away, which would make it trustworthy and accessible. But the waiter has not recorded a list of potential dishes, instead he has generated an intimate understanding of what *guides* the individual's choice, regardless of the circumstances (such as available ingredients or cooking methods) which makes an evaluation that bit more intricate.

I think, provided that the waiter is always available to the individual when called upon, the waiter could be considered as an extension of the individual's mind, even though they are performing a task that the individual would never be able to perform.

This is promising, however there is a recurring theme of a type of central controller that is required in order to justify a third-party decision that is as equally valid as an internal decision, which is essential for retaining a sense of self-hood when determining values. In the case of the waiter and the surgeon, taste acts as this central controller. The waiter's endeavour would be pointless without the subjective conscious experience of taste on behalf of the individual. This appears to mark the necessary component that distinguishes an extended mind from what I have been referring to as a 'distributed mind'. If the individual could not taste, then we could not say that the waiter's suggestions were extensions of the individual's mind. Of course, tastes can develop and change, and that would be fine in this example. Perhaps the waiter's accuracy may drop for a time, but it would increase given more observations, and we can presume that it would still remain better than the individual's decisions (provided it stayed above the level of an arbitrary guess, which, statistically speaking, is almost guaranteed).

This goes somewhat against the wording used in Clarke and Chalmer's paper, in which they state that the coupled system 'jointly govern[s] behaviour in the same sort of way that cognition usually does' (p. 8) I do not agree that this wording necessarily applies in cases where decision-making is extended to AI environments, because I do not agree that the extension should be able to govern behaviour if the concept of self is to be retained. For instance, the waiter may bring out a bad dish, and 'govern' behaviour in the sense that he made the surgeon eat a bad dish, but he does not 'govern' behaviour in the sense that he convinces the surgeon that the dish is nice. So yes, the AI environment may *provoke* certain actions (eating a certain meal), but I would not agree that it necessarily *governs* behaviour ('the behaviour' is the objective to eat a nice tasting meal, the waiter does not govern this, he only suggests ways to get to it).

These examples suggest that we should accept the theory of the extended mind in instances where the extension is performing some sort of decision-making process, provided it is governed by an internally generated system of values.

4.6 Where Should We be Focusing our Efforts?

Incalculable complexity for the biologically limited human brain is often understood to be the issue standing in the way of an individual determining their own values. Elements of this claim can be observed in empirical studies regarding will-power and decision fatigue (Baumeister and Tierney 2011; Hirshleifer et al. 2019; Pignatiello, Martin, and Hickman 2020), which is reportedly even more depleting in cases of making decisions concerning the self (Polman and Vohs 2016). This idea of overwhelming complexity is also explored philosophically in concepts such as Soren Kierkegaard's account of anxiety as the 'dizziness of freedom' (Kierkegaard and Hannay 2014) and Friedrich Nietzsche's 'labyrinth' in which the 'few [people that] are made for independence … [are] torn to pieces limb from limb by some cave-minotaur of conscience' (2014, 46).

If AI environments can fulfil certain necessary requirements, then they may become a tool that can facilitate a blended cognitive framework which enables individual moral responsibility and freedom. In order for this desirable future to manifest, AI environments will have to be constantly accessible to the individual (i.e. without down-time, prohibitive censorship, pay-walls, or even just physical separation or battery-life). Fortunately, tools are currently being produced that may enable a constant feed between a machine and human brain (known as Brain Machine Interfaces or 'BMIs'), such as Neuralink. More importantly, however, AI environments will be required to maintain a detailed encoding of the individual's moral 'tastes', in order to guide decision-making. This encoding may begin hazy and build up to an accurate model as we imagined in the case of the waiter and the surgeon, or perhaps it will be generated by neuro-scientific modelling like Sam Harris explores in *The Moral Landscape*. Whatever shape this guidance system may manifest in, the key hallmark is that it entails placing the individual at the centre of decision-making processes.

This proposed integration with AI environments provides an individual-centric and human-centric solution to issues that may arise from AI environments coming to resemble religions. When considering desirable futures, this option stands in opposition to the Singularity that is foretold and endorsed by techno-utopians. As we have briefly discussed previously, techno-utopians espouse a future where the individual becomes a part of a

universal hive-mind, which can be essentially considered a centralised value-system. This may, in fact, be a legitimate trajectory toward a utopian future, however I believe that AI environments would become the epitome of religions if this were to be the case. The promises of the Singularity are extremely vague, to the extent that it brings to mind concepts such as the 'Rapture' or return of Jesus Christ, where a currently inconceivable utopia will manifest on earth. This, as with any inconceivable scenario, requires a complete leap of faith on behalf of the individual. In order to commit to this scenario the individual would have to sacrifice themselves to the 'Data Religion', as Yuval Noah Harari refers to it in *Homo Deus* (2016, 428). The ontology of the 'Data Religion' is to interpret everything as data and information flowing. While I will not dwell on this scenario, as it leaves the scope of this research, it is noteworthy to explore the two outcomes that may arise, as they contrast so drastically with my proposal for human-centric AI environments. The first, more desirable outcome–from an individual point-of-view–is a technological utopia like the 'Fully Automated Luxury Communism' outlined by Aaron Bastani (2019), where the means of production are outsourced entirely to machines and humans are free to pursue leisure activities. The issue with this outcome is that it depends on a sharp distinction between means of production and leisure, and the assumption that human fulfilment should only occur on the side of leisure. This would require a leap of faith on behalf of the individual insofar as that they would sacrifice their job and trust that they could find a suitable means of self-determination in a world where their autonomy is restricted to choosing which leisurely activity to pursue. The second scenario is one where humans accept that they have fulfilled their cosmic use and retire into the shadows to allow whatever type of hive-mind/artificial consciousness that arises from the Singularity to flourish, in the same way that monkeys 'stepped aside' (albeit not consciously, nor peacefully) in order for homo sapiens to flourish. This appears to be the closest possible scenario in which AI environments replicate the mistakes that religions made in centralising value-systems. In this instance, these systems become dogmatic in their thinking, and begin to stagnate. We have seen that individuals do not thrive under those oppressive conditions, but rather only 'thrive' in the sense of propagating the existing dogma, which is commendable under the paradigm of the time and therefore difficult to identify as a force that denies freedom. At this point in time, to promote a hive-mind

singularity is to endorse a freedom-denying force, despite it looking so universally beneficial due to the current paradigm of Silicon Valley hype and the promise of AI.

My proposal places AI environments in stark contrast with what religions promoted, which was a centralised universal value-system, whereas what is required for self-determined values is a de-centralised universal value-system. As this research has explored, AI environments are on a current trajectory to replace religions, in every sense, including being the vessels for a centralised dogma that is used to shape the individuals that interact with it. If developments like the Neuralink become a reality, and individuals are biologically linked to an environment that modifies the individual to fit a value-system, then the ability to self-determine values will be irreversibly lost. On the other hand, if the necessary steps are taken to focus the available computing power and technological process on creating a robust de-centralised value-system, then AI environments could become the tool that finally enables humans to overcome themselves, where previously they had been restricted by biologically limited mental processing power.

Bibliography

Bastani, Aaron. 2019. *Fully Automated Luxury Communism: A Manifesto.* Verso Books.

Bastos, Marco, Dan Mercea, and Andrea Baronchelli. 2018. "The Geographic Embedding of Online Echo Chambers: Evidence from the Brexit Campaign." *PLOS ONE* 13 (11): e0206841. https://doi.org/10.1371/journal.pone.0206841.

Baumeister, Roy F, and John Tierney. 2011. *WillPower : Rediscovering the Greatest Human Strength*. The Penguin Press. https://www.researchgate.net/profile/Kathleen_Vohs/publication/318851043_Willpower_Choice_and_Self-Control/links/56f3c11a08ae38d7109bb704/Willpower-Choice-and-Self-Control.pdf.

Beebee, Helen. 2013. *Free Will - an Introduction.* Hampshire, UK: Palgrave Macmillan.

Bickle, John. 2020. "Multiple Realizability." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/multiple-realizability/.

Bodó, Balázs, Natali Helberger, Sarah Eskens, and Judith Möller. 2019. "Interested in Diversity." *Digital Journalism* 7 (2): 206–29. https://doi.org/10.1080/21670811.2018.1521292.

Bostrom, Nick. 2016. *Superintelligence.* Oxford University Press.

Boyer, Pascal. 2001. *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books. https://hdl-handle-net.ucc.idm.oclc.org/2027/heb.30744.

Boyer, Pascal, and Pierre Liénard. 2006. "Why Ritualized Behavior? Precaution Systems and Action Parsing in Developmental, Pathological and Cultural Rituals." *Behavioral and Brain Sciences* 29 (6): 595–613. https://doi.org/10.1017/S0140525X06009332.

Bret Weinstein on the Dawkins Debate. 2018. https://www.youtube.com/watch?v=rm8FksjlJtM.

Carr, Nicholas. 2008. “Is Google Making Us Stupid?” The Atlantic. July 1, 2008. https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/.

Chalmers, David J. 2016. “The Singularity: A Philosophical Analysis.” In *Science Fiction and Philosophy*, edited by Susan Schneider, 171–224. Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118922590.ch16.

Clark, Andy, and David Chalmers. 1998. “The Extended Mind.” *Analysis* 58 (1): 7–19.

Clavien, Christine, and Rebekka A. Klein. 2010. “EAGER FOR FAIRNESS OR FOR REVENGE? PSYCHOLOGICAL ALTRUISM IN ECONOMICS.” *Economics & Philosophy* 26 (3): 267–90. https://doi.org/10.1017/S0266267110000374.

Cohn, Samuel. 2019. *Race, Gender, And Discrimination At Work*. Routledge.

Dawkins, Richard. 2016. *The God Delusion*. London: Black Swan.

Dennett, Daniel C. 2007. *Breaking The Spell*. UK: Penguin Books.

———. 2014. *Intuition Pumps and Other Tools for Thinking*. London: Penguin.

Domingos, Pedro. 2017. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. UK: Penguin Books.

Douglass, Frederick, Stanley Applebaum, and Philip Smith. 1995. *Narrative of the Life of Frederick Douglass, an American Slave*. Mineola, NY: Dover Publications Inc.

Durant, Will. 1927. *The Story of Philosophy*. New York: Simon and Schuster.

Esteban, A. 1999. “A Neurophysiological Approach to Brainstem Reflexes. Blink Reflex.” *Neurophysiologie Clinique/Clinical Neurophysiology* 29 (1): 7–38. https://doi.org/10.1016/S0987-7053(99)80039-2.

Evans, Jules. 2018. *The Art of Losing Control: A Philosopher’s Search for Ecstatic Experience*.

Fan, Rui, Ke Xu, and Jichang Zhao. 2016. “Higher Contagion and Weaker Ties Mean Anger Spreads Faster than Joy in Social Media.” *ArXiv:1608.03656 [Physics]*, December. http://arxiv.org/abs/1608.03656.

Fifel, Karim. 2018. “Readiness Potential and Neuronal Determinism: New Insights on Libet Experiment.” *The Journal of Neuroscience* 38 (4): 784–86. https://doi.org/10.1523/JNEUROSCI.3136-17.2017.

Frankfurt, Harry G. 1969. “Alternate Possibilities and Moral Responsibility.” *The Journal of Philosophy* 66 (23): 829–39. https://doi.org/10.2307/2023833.

Fridman, Lex. 2020. “#103 - Ben Goertzel: Artificial General Intelligence | MIT | Artificial Intelligence Podcast.” *Lex Fridman* (blog). June 22, 2020. https://lexfridman.com/ben-goertzel/.

Frischmann, Brett, and Evan Selinger. 2018. *Re-Engineering Humanity*. UK: Cambridge University Press.

Garde, M. M., and A. Cowey. 2000. “‘Deaf Hearing’: Unacknowledged Detection of Auditory Stimuli in a Patient with Cerebral Deafness.” *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior* 36 (1): 71–80. https://doi.org/10.1016/s0010-9452(08)70837-2.

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. “Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship.” In *Proceedings of the 2018 World Wide Web Conference*, 913–22. WWW ’18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186139.

Hájek, Alan. 2018. “Pascal’s Wager.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/pascal-wager/.

Harari, Yuval Noah. 2016. *Homo Deus: A Brief History of Tomorrow*. London: Vintage Books.

Harris, Sam. 2012. *The Moral Landscape: How Science Can Determine Human Values*. London: Transworld Publishers.

Hirshleifer, David, Yaron Levi, Ben Lourie, and Siew Hong Teoh. 2019. "Decision Fatigue and Heuristic Analyst Forecasts." *Journal of Financial Economics* 133 (1): 83–98. https://doi.org/10.1016/j.jfineco.2019.01.005.

Howell, Robert J. 2014. "Google Morals, Virtue, and the Asymmetry of Deference." *Noûs* 48 (3): 389–415. https://doi.org/10.1111/j.1468-0068.2012.00873.x.

https://kurzweilai.net. n.d. "The Hivemind Singularity « Kurzweil." Accessed September 6, 2020. https://www.kurzweilai.net/the-hivemind-singularity.

Jung, C. G, W. S Dell, and Cary F Baynes. 2001. *Modern man in search of a soul*. London: Routledge.

Kierkegaard, Søren, and Alastair Hannay. 2014. *The Concept of Anxiety: A Simple Psychologically Oriented Deliberation in View of the Dogmatic Problem of Hereditary Sin*. New York; London: W. W. Norton.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences* 111 (24): 8788–90. https://doi.org/10.1073/pnas.1320040111.

Kuss, Daria J., and Mark D. Griffiths. 2017. "Social Networking Sites and Addiction: Ten Lessons Learned." *International Journal of Environmental Research and Public Health* 14 (3): 311. https://doi.org/10.3390/ijerph14030311.

La Boétie, Étienne de. 1975. *THE POLITICS OF OBEDIENCE: The Discourse of Voluntary Servitude*. Translated by Harry Kurz. Quebec, Canada: Black Rose Books.

Lanier, Jaron. 2011. *You Are Not a Gadget*. New York: Vintage Books.

Leeuwen, Neil Van, and Michiel van Elk. 2019. "Seeking the Supernatural: The Interactive Religious Experience Model." *Religion, Brain & Behavior* 9 (3): 221–51. https://doi.org/10.1080/2153599X.2018.1453529.

Levin, Janet. 2018. “Functionalism.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/functionalism/.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. “Algorithm Appreciation: People Prefer Algorithmic to Human Judgment.” *Organizational Behavior and Human Decision Processes* 151 (March): 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005.

Lopez, Donald S. 2004. *Buddhist scriptures*.

Magoulick, Mary. 2004. “What Is Myth?” 2004. https://faculty.gcsu.edu/custom-website/mary-magoulick/defmyth.htm.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. London: John Murray Publishers.

Müller, Vincent C., and Nick Bostrom. 2016. “Future Progress in Artificial Intelligence: A Survey of Expert Opinion.” In *Fundamental Issues of Artificial Intelligence*, edited by Vincent C. Müller, 555–72. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_33.

Neumann, Erich, C. G Jung, and Richard Francis Carrington Hull. 2014. *The Origins and History of Consciousness*. Princeton; Woodstock: Princeton University Press.

Nietzsche, Friedrich Wilhelm, R. J Hollingdale, and Michael Tanner. 2014. *Beyond good and evil: prelude to a philosophy of the future*.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5.

Peterson, Jordan B. 1999. *Maps of Meaning: The Architecture of Belief*. New York: Routledge.

Pignatiello, Grant A, Richard J Martin, and Ronald L Hickman. 2020. “Decision Fatigue: A Conceptual Analysis.” *Journal of Health Psychology* 25 (1): 123–35. https://doi.org/10.1177/1359105318763510.

Polman, Evan, and Kathleen D. Vohs. 2016. “Decision Fatigue, Choosing for Others, and Self-Construal.” *Social Psychological and Personality Science* 7 (5): 471–78. https://doi.org/10.1177/1948550616639648.

Poundstone, William. 2017. *Head in the Cloud : Dispatches from a Post-Fact World*. Mass Market edition. London: Oneworld Publications.

Richard Dawkins & Bret Weinstein - Evolution. n.d. Accessed September 30, 2020. https://www.youtube.com/watch?v=hYzU-DoEV6k.

Rodríguez-García, Antonio-Manuel, Antonio-José Moreno-Guerrero, and Jesús López Belmonte. 2020. “Nomophobia: An Individual’s Growing Fear of Being without a Smartphone—A Systematic Literature Review.” *International Journal of Environmental Research and Public Health* 17 (2): 580. https://doi.org/10.3390/ijerph17020580.

Russell, Bertrand. 2010. *History of Western Philosophy*. [New ed.], Repr. London: Routledge.

Russell, Stuart J. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. London: Allen Lane/Penguin Random House.

Schmidtz, David. 2019. “Friedrich Hayek.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/friedrich-hayek/.

Schultze-Kraft, Matthias, Daniel Birman, Marco Rusconi, Carsten Allefeld, Kai Görgen, Sven Dähne, Benjamin Blankertz, and John-Dylan Haynes. 2016. “The Point of No Return in Vetoing Self-Initiated Movements.” *Proceedings of the National Academy of Sciences* 113 (4): 1080–85. https://doi.org/10.1073/pnas.1513569112.

Sinding Jensen, Jeppe. 2009. “Religion as the Unintended Product of Brain Functions.” In *Contemporary Theories of Religion*, edited by Michael Stausberg, 129–55. London: Routledge.

St. Marie, Raymond, and Kellie S. Talebkhah. 2018. “Neurological Evidence of a Mind-Body Connection: Mindfulness and Pain Control.” *American Journal of Psychiatry Residents’ Journal* 13 (4): 2–5. https://doi.org/10.1176/appi.ajp-rj.2018.130401.

Stark, Rodney, and Roger Finke. 2007. *Acts of Faith: Explaining the Human Side of Religion*. Berkeley: Univ. of California Press.

Tegmark, Max. 2017. *Life 3.0*. 1st ed. USA: Penguin Books.

Tenbarge, Kat. n.d. "Justin Bieber Is Desperate to Get His New Song 'Yummy' to No. 1 and Is Asking Fans to Manipulate Streaming Services to Do It." Insider. Accessed September 23, 2020. https://www.insider.com/justin-bieber-get-yummy-to-no-1-despite-lackluster-response-2020-1.

Tétreault, Pascal, Ali Mansour, Etienne Vachon-Presseau, Thomas J. Schnitzer, A. Vania Apkarian, and Marwan N. Baliki. 2016. "Brain Connectivity Predicts Placebo Response across Chronic Pain Clinical Trials." *PLOS Biology* 14 (10): e1002570. https://doi.org/10.1371/journal.pbio.1002570.

Wallace, David Foster. 1997. *Infinite Jest*. London: Abacus.

Watts, Alan, and Al Chung-liang Huang. 1975. *Tao: The Watercourse Way*. New York: Pantheon Books.

"What Happened to the Future?" n.d. Founders Fund. Accessed September 30, 2020. https://foundersfund.com/the-future/.

Whitehouse, Harvey, Pieter François, Patrick E. Savage, Thomas E. Currie, Kevin C. Feeney, Enrico Cioni, Rosalind Purcell, et al. 2019. "Complex Societies Precede Moralizing Gods throughout World History." *Nature* 568 (7751): 226–29. https://doi.org/10.1038/s41586-019-1043-4.

Wylie, Christopher. 2019. *Mindf*ck: Inside Cambridge Analytica's Plot to Break the World*.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*.