How can Factors Underlying Human Preferences lead to Methods of Formal Characterizations towards Developing Safe Artificial General Intelligence?

July 2020

Glasgow Student Number: 2389462A

Dublin City University Number: 18114075

Charles University Number: 55140763

Rana Ghalib Sami Al-Nusair

Presented in partial fulfilment of the requirements for the Degree of

International Master in Security, Intelligence and Strategic Studies

Word Count: 23,379

Supervisor: Petr Špelda

Date of Submission: July 26th 2020

International Master in Security, Intelligence and Strategic Studies 2018/2020

Dissertation Archive Permission Form

I give the University of Glasgow and Charles University permission to archive an e-copy of my Master dissertation in a publicly available folder and to use it for educational purposes in the future.

Student Name: RANA GHALIB SAMI AL-NUSAIR

Student Number: 2389462A

Student Signature:  **Date:** July/26/2020

Acknowledgements

I would like to give thanks to the following people, whom have helped me tremendously throughout the process of writing this paper:

Thank you to my supervisor, Petr Špelda, who has always been supportive of me, and has truly expanded my knowledge horizon and deepened my passion for the work I do. I could not have gone this far without your continuous reassurance and your faith in me.

Thank you to my parents, who have always believed in me and have helped me in all possible ways, and always supported me through thick and thin. I would not be here without you.

Thank you to my partner, Shorouq Zahra, for your continuous encouragement, your valuable critique, and your constant intellectual engagement with me. I have always appreciated your honesty and objectivity.

Thank you to Abu Saqir and Ranchwood, for providing me with ease of mind, reassurance, and comfort throughout the Covid-19 situation during the time of writing this paper. Reaching this stage in these harsh circumstances would have been much more difficult without your support.

Finally, I have the deepest gratitude for Terence Burnham. Thank you for taking up the time to engage with my work and provide me with your valuable insight and critique.

Abstract

This research aims to investigate the Artificial Intelligence (AI) value alignment problem, which refers to the challenge in developing a safe and reliable AI that can achieve our goals and adhere to our values as we intend it to do. A misaligned AI, especially one which transcends all domains of cognitive abilities and has acquired vast computational powers, will be nearly impossible to manage and it will threaten our security. Research addressing this problem is now focused on understanding how to develop AI that can reliably infer our values from our preferences. Thus, preferences are the primary conceptual unit of analysis to the AI value alignment problem. This paper investigates our preferences and seeks to shed light on the issue of obtaining a formal truth that is fundamentally constitutive of our preferences, for the aim of using said formal truth to create a value aligned AI. To do this, this paper gathers data from economics, biological evolution, and neurocognitive studies to bridge the current gaps on the conceptual problem of preferences. The paper concludes with presenting a new kind of security dilemma which stems from the notion of combining a general theoretical framework that fully captures our preferences with the crucial element of uncertainty in AI, effectively showcasing how the two ideas contradict one another. The paper suggests that future research in AI must focus on developing a framework which can reconcile formal truth about our preferences with the key element of uncertainty. This finding serves to show that the problem of AI control is potentially much more difficult to solve than we had initially anticipated, which implies an even greater urgency to support AI value alignment research and encourage more people to work on the problem.

Table of Contents

Acknowledgements

Abstract

Table of Contents

Chapter One: Literature Review 1

1.1 How Can AI Become Dangerous 3

1.2 Why is AI Research Crucial to Security Studies? 5

1.3 How Will AI Affect the Future of Governance? 7

1.4 How Can ASI Come About? 8

1.5 Rationale and Area of Intervention 9

Chapter Two: Research Design 15

2.1 Methodology 16

2.2 Scope and Limitations 18

Chapter Three: The Evolution of Preferences 19

3.1 Preferences and Maximization 24

3.2 Expected Utility, Trade Offs, and Discounting the Future 29

3.3 Risk Preferences and Uncertainty 36

Chapter Four: The Neurocognitive Science of Preferences 42

4.1 Preferences: AI Modelling, Where We Are Now 43

4.2 Preferences: Memories and Intentions 49

4.3 Preferences: Memories and Choice Sets 52

4.4 Preferences: Brain Modularity 60

Chapter Five: Analysis 65

5.1 Discussion 65

5.2 Conclusion 74

References 77

Chapter One: Literature Review

This research aims to investigate current work on the Artificial Intelligence (AI) value alignment problem (I.e. the AI control problem). The AI value alignment problem refers to the challenge in developing a safe and reliable AI that seeks to achieve our goals and values in the manner in which we intend. If AI is aligned with our goals and values then it would not pose any significant threat to our security and wellbeing. Research addressing this problem is fairly recent, and the control problem has proven to be a difficult one to solve. The main difficulty here emerges from the notion of there being no general theoretical framework that can accurately capture human values and preferences, not to mention one which could be modelled mathematically.

Proponents of the AI value alignment warn of the dangers of developing AI that is capable of indefinite self-improvement (I.e. recursive self-improvement) but is not aligned with our values and does not do what we intend it to do for us. An unaligned AI, especially one which has tremendous capabilities and global access, can produce numerous security threats to our survival, some of which we can foresee (e.g. control of nuclear weapons), but most of which we cannot predict. The prospect of not being able to predict the actions of an unaligned AI entails that we will be left far behind in terms of learning how to manage, contain, or even shut it down (Omohundro, 2008; Bostrom, 2014a; Yudkowsky, 2016a; Tegmark, 2017; Russell, 2019).

This paper will focus on the concept of preferences and will tackle them through a transdisciplinary lens, for the aim of extrapolating some fundamental key elements which may constitute our preferences. Understanding the concept of preferences is crucial towards the development of a reliably beneficial AI that is aligned with our values. The reason for this is

because there appears to be no way to directly encode our values into a system, given that our values are arbitrary, unstable, and often conflicting (Tegmark, 2017; Chalmers, 2010). Research in AI has now turned to methods which seek to indirectly infer our values from our preferences as revealed through our behaviours. Russell (2019; 176) argues that a potentially beneficial AI, in contrast to the outdated standard model, can gradually learn our values if we can obscure its reward function and force it to try to infer our values instead by studying our preferences and have it refer back to us for instructions when it is faced with *uncertainty*. This new model of AI can potentially avoid some of the pitfalls that are present in the standard model, which has been dominating the field for many years now. Thus, I will focus on preferences by investigating their nature and origin to inform the current research on AI alignment with the use of conceptual case study analysis.

Artificial General Intelligence (AGI) refers to the idea of a machine becoming as intelligent as a human in a broad range of cognitive abilities to the extent where we are unable to distinguish this machine from a human being. Artificial Superintelligence (ASI) refers to the point where AGI far exceeds the best human to have ever lived in all domains of cognitive abilities. For an AGI to become ASI, it would first have to go through the process of *intelligence explosion* where an AGI that can recursively self-improve will quickly and vastly become superior to us in a wide range of capabilities by virtue of continuous acceleration in its processing power and speed, which would come parallel with its increase in intelligence (Good, 1965, 31-88; Sotala, 2018a, 320-323). ASI represents an emerging research discipline in safety and security, one that is gaining more traction as we observe the speed at which the current technological advancements are taking place over this century.

Given the idea of ASI being an emergent security problem, the threats arising from this future hypothetical instance come from various angles, some of which are not immediately obvious. An ASI, by virtue of being far more intelligent and capable than us, has the potential to greatly disrupt our landscape and irreversibly incapacitate our growth in ways which we cannot foresee and much less control, if its motivations are not in alignment with ours. For an ASI to be reliably beneficial to us, it must truly understand what we value. Therefore, the primary risk of an ASI arises from the conceptual problems regarding the human condition, as opposed to the ones which relate to the direct programming of this agent (Yudkowsky, 2016a: 1-19).

1.1 How Can AI Become Dangerous?

AI can pose numerous threats to our security. Any sufficiently intelligent AI that has been supplemented with a direct and specific objective function will seek to maximize it. For example, assume you have a helper robot and you ask this robot to fetch you a book. The robot will calculate the fastest and most efficient route to achieving this objective. In order to succeed at this, the robot creates sub-goals (I.e. goals that are not directly encoded into it, but are crucial to achieving the primary goal that has been directly encoded). The robot, as it attempts to achieve the goal of fetching you a book, may knock over things in its environment (e.g. other books, a vase, kitchenware). The goal was to fetch the book, anything else that may affect the speed and efficiency at which the robot can achieve this objective will be rendered irrelevant, or rather, an obstacle. Additionally, scouring through an ocean of possibilities of either succeeding or failing at the objective, the robot calculates that the possibility of being turned off would prevent it from succeeding. You cannot fetch the book if you are turned off. Thus, the self-perseverance subgoal emerges. The above-mentioned example is overly simplified. However, the self-perseverance subgoal can become dangerous when taken in different situations and more dynamic contexts;

an AI that cannot be shut down is an AI that cannot be controlled. If AI obtains control of nuclear warheads and it is not in full alignment with our intentions, then we must be able to shut it down when necessary.

Omohundro (2008) discusses other emerging subgoals, arguing that these subgoals are bound to be deemed as necessary for an AI to successfully achieve any and all kinds of objectives that we may supply the AI with. Unless we very carefully design this AI, these subgoals will invariably emerge. Another equally important emerging subgoal is that of resource acquisition. This means that when supplying the AI with an objective, the AI will need to acquire as many resources as possible in order for it to calculate the optimal decision to employ in achieving the objective. Resource acquisition can even become unbounded, where an AI with a certain objective will need unbounded resources in order to maximize its objective, and this may include converting the entirety of our landscape into a global industrial platform if it were an ASI, effectively disregarding all humans who live on this landscape (Bostrom, 2014b; 111-114). Another dangerous subgoal is that of utility or goal perseverance. As the AI keeps improving upon its own code, it will need to ensure that its initial goal remains intact and does not change throughout the process of self-recursive improvement. Thus, it develops the subgoal of goal perseverance. This is a dangerous subgoal as it entails that as the AI becomes sufficiently more powerful, we will be left incapacitated in terms of reversing or adjusting the AIs goal structure if this structure does not align with our values in the later development phases of the AI, as the AI will be far better than us at resisting any attempt from our side in revising its code (Omohundro, 2008).

1.2 Why is AI Research Crucial to Security Studies?

AI would be relevant to every aspect of our lives, but more importantly, it would become solely and directly responsible for our collective human security as it progresses into AGI, and ASI soon after. For example, an unaligned ASI with global access would be able to obtain nuclear launch codes, destroy the global economy, or engineer chemical and biological weapons far better than any human ever could (Bostrom, 2018b: 152). In Superintelligence: Paths, Dangers and Strategies, Bostrom (2014a: 96-99) expands on such scenarios. Take the example of implementing a direct value into the AI, such as happiness. AI researchers can supplement the AI with the goal 'make us happy'. Such a goal would lead to what Bostrom (2014) calls *perverse instantiation*, meaning that as the AI attempts to achieve this goal, it inadvertently ends up achieving something completely different from what we had intended. In this example, the AI might do something like plant electrodes into the pleasure centers of our brain, effectively making us 'happy', even though this is not what we were truly aiming for. These examples exclude the ones which we would not be able to foresee due to an ASI being capable of engineering evermore cost-effective and less direct yet nonetheless just as or even more catastrophic events. On the other hand, an ASI that is aligned with our values can end poverty, permanently uplift the global economy, cure diseases, and overall greatly enhance the ways in which we live our lives (Sotala, 2017b: 501-510). ASI can either effectively resolve all of the security dilemmas which we consistently face throughout our lives, or it can become the most prominent one in all of human history. As many researchers in AI and other fields believe, we are likely headed into the future where ASI emerges, assuming there occurs no major obstacle that would slow down, stop, or outlaw research and development in the field (such as in the case of a nuclear disaster or global war). From this notion follows the concept of AI race in a highly competitive global market, where

an AI race could come at the cost of sparing resources and time to first finding a solution to the value alignment problem for the advantage of having 'won' the race.

The value alignment problem currently contains a significant research gap in AI that mainly stems from our inability to formally define a coherent decision-making framework which relies on having first clearly identified the preceding concepts that are crucial to developing this framework (e.g. values, preferences, wants and desires). If ASI emerges before having solved the value alignment problem first, there will be no shutdown button that we can press to go back to how things were before the instance of emergence (Russell, 2019: 115-126). The prospect of irreversibility is one of the many factors which serve to showcase that the problem of alignment is one that must be solved before the instance of emergence, and that we only have one chance of getting it 'right'.

We do not know when an intelligence explosion will happen, nor whether we will be able to accurately anticipate this explosion and become aware of it before or even while it is happening. An important aspect relating to national and international security is whether governments and various authorities will be able to anticipate when ASI will emerge. The idea of ASI has been widely portrayed in media and film, albeit as a fanciful idea that does not accurately reflect the state of the subject matter and overshadows its true dangers. Thus, ASI is not taken as seriously by some authorities and various segments of the public and academia. In 1933, Ernest Rutherford said that nuclear energy was "moonshine". Twenty-four hours later, Leo Szilard's proved him wrong by inventing the nuclear chain reaction. In relation to this, Tegmark (2017) says:

> The most extreme form of this myth is that superhuman AGI will never arrive because it's physically impossible. However, physicists know that a brain consists of quarks and electrons

> arranged to act as a powerful computer, and that there's no law of physics preventing us from building even more intelligent quark blobs. (Ch. 1, *Timeline Myths*, para. 2)

Here, Tegmark (2017) justified the argument of creating AGI, and later ASI, by explaining that the brain is computational or mechanical, meaning that its underlying processes can be unlocked, replicated, and soon afterwards become fully programmable. In other words, our brains can be understood mathematically, which means that they can be formally computed in an artificial system. Once this is achieved, the artificial system will be able to drastically improve upon its own brain and become superintelligent.

1.3 How will AI Affect the Future of Governance?

Once and if progress in AI begins hinting at an impending intelligence explosion, governments and authorities will begin to closely monitor teams and individuals working on AI. This entails heavy surveillance, regulations, and potentially even outlawing work in the field or nationalizing it in some countries (Bostrom, 2014b: 246-249). This has some serious implications. For example, if governments and other state-actors begin to see signs of an impending intelligence explosion, this will increase their competitiveness and rush their work towards gaining a first mover advantage against other AI teams. This AI rush would come at the expense of AI safety. If work on the AI value alignment or AI control problem is not prioritized, then this would speed up the race into an unpredictable future that will most likely be dictated by an ASI which we cannot change or control.

1.4 How Can ASI Come About?

There are numerous paths from which ASI can emerge: a) artificial intelligence/machine learning, b) brain emulation, c) cognitive enhancement, d) networks & organizations, e) genetic enhancement (Bostrom, 2014c: 64-65). Take brain emulation, for example. One way of attaining AGI might be through replicating the dynamic processes and patterns of the brain. With technological advancement, we are now learning more and more about how our brains function, and it seems that they are -at least to a considerable extent- computational or mechanical. If our brains are indeed computational, then replication is theoretically possible. Another method is artificial intelligence, which entails writing up the AI program from scratch. This method can directly lead to AI, however, it is lengthy and riddled with problems, given that it would require having first solved the value alignment problem and other problems regarding decision making and performing abstractions. Machine learning, on the other hand, is a bit more promising than this. Machine learning is a subset of artificial intelligence, which seeks to develop systems that are able to learn on their own, without the need for constant human supervision and human supplementation of data. With machine learning, the system can come to improve upon itself by itself. Cognitive enhancement considers the engineering of ever more powerful substances that can enhance our intelligence way beyond the current threshold. Genetic enhancement is similar to this, as it involves modifying our very genetic composition to select for more competent offspring that would eventually move up the hierarchy of intelligence (Brin, 2018, 114-125; Chalmers, 2010; 7-65). The likelihood of ASI emerging from at least one or a combination of multiple paths increases. As Nick Bostrom and others believe, even if some of those methods are not immediately feasible towards reaching above human-level intelligence, there is at least one or a combination of two or more methods which will lead to a path of emergence (Bostrom, 2014d: 22-49). However,

it is not immediately obvious when ASI might emerge. Some researchers believe that it will be in the middle of this century, while others maintain that it will be much longer than that (Russell, 2019, 78-79; Tegmark, 2017).

1.5 Rationale and Area of Intervention

This research aims to primarily investigate the underpinnings of our preferences in order to reach a better understanding of how an AI can infer our values from said preferences. However, before I delve into this, I will provide a brief overview of where we are now in terms of developing safe and reliable AGI, which may later converge onto an ASI.

In 2019, DeepMind's Alphastar has reached Grandmaster level in the online multiplayer real-time strategy game StarCraft, which is deemed to be the most difficult game of its genre. Alphastar reached a 99.8%-win rate against human players, and it managed to do so with a combination of supervised deep neural networks and reinforcement learning algorithms (The Alphastar Team, 2019). Before Alphastar, DeepMind's OpenAI Five beat the top professional human team in the online multiplayer game of Dota 2, and it had learned to do so by playing over 10,000 games against itself (OpenAI, 2019). Within the domain of real-time strategy gaming, DeepMind's Alphastar and OpenAI Five are considered superhuman, although they are still considered narrow AI in the sense that their current level of intelligence cannot extend to the broader domains of cognitive abilities.

Regarding deep learning (which is a subset of machine learning), artificial neural networks that are inspired by the human brain are able to take in large sets of structured and unstructured data, and with a combination of supervised and unsupervised learning, are then able to make sense of the data and solve complex problems. Regarding Neural Machine Translation (NMT),

Facebook achieved first place in 2019 in the annual news translation competition, which included translating text from English to German, English to Russian, and vice versa. The Fourth Conference on Machine Translation (WMT) dubbed Facebook's AI superhuman, as the human evaluators for the AI's tasks preferred the AIs translation over expert human translation (Ng *et. al.*, 2019). Meanwhile, the tech giant Google had created the largest NMT system as they managed to incorporate an even larger amount of training data and languages than any other tech giant to date (Wu *et. al.*, 2016). However, Google's AI is currently trained on only 103 languages, and if it were to incorporate all languages in its training data then that would pose a significant scaling challenge which would result in accuracy reduction.

In terms of general-purpose AI technology, in November 2019, OpenAI released the GPT-2, which is an AI that is capable of generating text when given sample texts. The GPT-2 was trained on 1.5 billion parameters, effectively making it quite impressive in terms of producing coherent and logical text that can be indistinguishable from text that is produced by a human. At first, OpenAI refused to release the source code because if it were to fall into the wrong hands, malign actors would use it in producing highly intelligent computational propaganda (I.e. fake news). In May 2020, OpenAI announced the GPT-3, which has been trained on *175 billion parameters* (OpenAI website, 2020). This represents a huge leap in a short period of time since the GPT-2.

Toshiba's Simulated Bifurcation Algorithm from 2019 represents a breakthrough in combinatorial optimization (a subset of mathematical optimization relevant to various fields of work). This algorithm enables the user to sift through a large number of combinatorial patterns to be able to extract the optimal solution for a given situation or problem. This algorithm can be highly beneficial in a range of applications such as financial portfolios, identifying efficient delivery routes, and drug development (Toshiba Corporation, 2019).

Decision Theory is a crucial aspect to the value alignment problem. To achieve a reliably beneficial AGI, there needs to be a good decision theory which we can supplement the agent with, one which would be reliable, consistent, and ensures that the agent does not deviate from what we intend it to do. Causal Decision Theory (CDT), Timeless Decision Theory (TDT) and Evidential Decision Theory (EDT) have not been successful in producing the kind of reliably beneficial intelligent performance that we aim for (Demski, 2018; Demski and Garrabrant, 2019). However, there have been emerging decision theories that show some improvement in the agent's performance. One is Functional Decision Theory (FDT), which has achieved better results than CDT and EDT on various decision-making dilemmas, such as Newcomb's paradox, the smoking lesion dilemma, and has been more successful than both in Parfit's hitchhiker problem (Yudkowsky and Soares, 2018). However, we are still far from developing a more general and coherent decision theory which can be used to create a provably reliable and beneficial AGI. The crucial problem in all of those proposed decision models is that none of them are currently applicable in real-world scenarios, thus making extrapolation of a general theoretical framework for a reliable decision theory quite challenging.

Developing a reliable decision theory might be less complicated than trying to directly encode final values in an intelligent agent. However, developing an agent with the wrong decision theory might result in a chain of erroneous decisions which can lead to a catastrophe due to the agents' initial false inferences. There is still much work to be done in terms of understanding how to create an agent that would not maximize some arbitrary goal which can be disastrous to us. The main problem lies in understanding how to formulate a theory which can reliably and beneficially capture our goals, and function well under *uncertainty*, with or without an agreed upon code of ethics which can possibly be applicable to all of us. This topic has been the subject of endless

debate between academics of different fields, and it serves to show how difficult it will be to solve the value alignment problem, in addition to the urgency of prioritizing work on it.

Considering the above-mentioned advancements, and many more that have come through within this decade alone, AI is still narrow as there is no specific algorithm yet which can generalize and perform abstractions in the same manner that we do. The recent developments look promising, but nonetheless remain limited in terms of reaching AGI, and later ASI, any time soon. The gap between what we have now and AGI is due in part to limitations in software (e.g. better algorithms), hardware (e.g. higher processing power and speed) and theory (e.g. lacking a reliable decision theory model and a model that accurately reflects our values and preferences). The relationship between the decision model and preferences is important. A sufficiently intelligent AI would still not be capable of navigating our cognitive myopia and inconsistencies, reliably inferring our preferences, without a solid decision theory model to compliment the methods in which this AI can infer our values.

Considering the broad scope of this paper, the primary area of intervention here will be on the concept of preferences. The aim of this research is to investigate a more comprehensive theoretical framework that can potentially shed light on formal characteristics for any fundamental underpinnings of human preference structuring, and how those preferences explain past decisions and guide future ones. If AI is to maximize our human preferences, then it will be forced to conduct trade-offs while taking into consideration our dynamic and inconsistent nature. Additionally, it will also be important to learn whether our preferences can potentially converge in the future on a global scale (I.e. become more ‘rational’ and consistent) or whether they will remain as divergent as they are now, or even diverge further due to the introduction of AGI.

It sounds reasonable that, if we want to understand what underpins our preferences in hope of identifying potential formal markers which constitute them, then we would have to go back to the beginning by investigating the biological and evolutionary origins of our preferences. The biological evolutionary framework can be beneficial in two particular ways. Firstly, it can aid our understanding of preferences by providing a more comprehensive tractability throughout our history. Biological evolution can help us track the inconsistencies and irrationality which presently underlie our preferences, by comparing our current preference orderings to those which our ancestors displayed (Burnham and Phelan, 2019; Irons, 1998; Hintze et. al., 2015). Secondly, a biological evolutionary framework may provide us with a more general understanding of our preferences as a species, rather than as separate individuals. This insight can particularly help us devise better predictions about the direction of our preferences in the future, by hypothesizing whether our preferences can converge on a collective scale, or whether they will continue diverging and become more inconsistent and irrational. This kind of insight might be crucial to AI researchers, given that AI will most definitely need to conduct trade-offs between large sums of different people across various nations. If our preferences continue diverging, then AI trade-offs might create security problems and further strengthen existing ones.

This paper will also incorporate theory from mainstream economics and behavioural economics, given that the study of human preferences proliferated from those disciplines and had become dominated by them. Mainstream and behavioural economics have long dominated the concept of preferences, human behaviour, intention, and value (Samuelson, 1948; Kahneman and Tversky, 1979). AI research derives theory from economics and behavioural economics and incorporates their findings into the current research on AI value alignment. Thus, it is necessary to investigate what mainstream and behavioural economics theory can inform us about

preferences. It might be beneficial if biological evolution can be integrated as a tool that can help bridge the gaps between the differences and similarities arising from mainstream and behavioural economics to inform the work on AI for the aim of providing formal markers.

Lastly, in order to connect all of the data into a coherent conceptual framework and showcase the present gaps in the literature of preferences, it is necessary to refer to the most recent findings from the neurocognitive sciences on our preferences, to provide a modern empirical basis for explanation and prediction. Incorporating findings from the neurocognitive sciences can be quite beneficial for us in terms of gaining a better overall understanding of our preferences in a more formal fashion. For example, neurocognitive science can help us map out specific regions within the brain which could be dominating the processes underlying preference formulation and adaptation (Dolan and Sharot, 2011; Naselaris et. al., 2018). If there exists a dominating region or subsystem within the brain that is directly in charge of preferences, then perhaps the functioning of this subsystem can be simulated through machine learning or deep learning techniques. Simulating the functioning of a brain region can help us build AI that is better able to understand and truly capture our preferences.

Chapter Two: Research Design

This research aims to address a transdisciplinary conceptual problem relating to the artificial intelligence (AI) value alignment problem, which is an emerging area of security studies. The AI value alignment refers to the challenge that multidisciplinary researchers have been facing in building an AI that can reliably infer and realize our human values on a collective, rather than on an individual scale. The AI value alignment is an important security problem which concerns not only individuals from within the field of AI, societies or groups of people, but rather it is one that concerns the safety and security of all of us. Developing an AI that has vast computational powers, resources and a global reach, will pose an existential threat to us if it is not aligned with our goals and values.

The conceptual problem I will address here is that of human preferences, how they emerge, adapt, and cohere or diverge. The study of human preferences has been recently deemed to be crucial in AI research on the value alignment problem. Past research on AI has been focused on building a general theoretical framework from which we can directly extrapolate and then encode our goals and values into the AI program. However, doing this has proven to be exceptionally difficult, if not impossible. The notion of human values has been fiercely debated by scholars coming from different fields of research for a very long time, yet we never reached a consensus on the subject. Instead, given the challenge of understanding and trying to directly encode our values, AI research has now shifted to devising algorithmic methods for indirectly inferring our values from our preferences through observing our behaviour.

The study of human preferences spans over more than a century, with it becoming the pinnacle of standard economics theory, and later, behavioural economics. The two fields have largely influenced the ways in which AI research is now viewing our preferences, however, they are not without some shortcomings and ambiguities. These shortcomings can be reconciled through the integration of further research on human preferences that primarily stems from biological evolution and neurocognitive studies. Thus, this research incorporates a transdisciplinary approach to investigate the conceptual case study of preferences, with the aim of bridging the gap between the various explanations of which those different disciplines give to the concept of human preferences. This can be done by attempting to find common themes and patterns across the board, ones which can be identified as formal markers for our preferences, which can be reliably tractable. If we are able to better understand the underpinnings of human preferences, with the aid of more empirical research on the subject, then we may be able to finally extrapolate a general theoretical framework that fully captures our preferences. In turn, this framework can be more effective for AI value alignment research, in terms of building a reliably beneficial AI that accurately infer our goals and values from said framework, and effectively mitigate the potential risks which will arise from this AI if this problem is left unsolved.

2.1 Methodology

To investigate the underpinnings of human preferences in order to explore the depth of the current gaps and shortcomings in AI value alignment research, I will utilize the method of transdisciplinary conceptual case study to analyze the process of preference formulation and adaptation as its primary unit. A persistent problem in research on AI inferring values from human preferences is that the current techniques and AI models cannot be reliably transferred to real-world context sensitive situations, which makes them ineffective at predicting reliability

of future beneficial cooperation between AI and human beings. Thus, I justify the conceptual case study analysis method considering that, in addition to the boundaries between notions of preferences and values in AI value alignment research being not clearly defined, the notion of preferences itself is not clearly defined either. Additionally, the current understanding of human preferences in AI research seems to lack a comprehensive and more formal analysis which can be obtained from outside the lens of mainstream economics and behavioural economics.

The conceptual case study analysis on preferences is bounded by three key propositions: our preferences are stable; our preferences maximize our objectives; and that we succeed in maximization. These three propositions will be investigated through a conceptual framework that combines literary sources from multiple disciplines: economics; behavioural economics; biological evolution; and neurocognitive sciences. The structure of the upcoming chapters in this paper will be as follows:

Chapter three starts with the introduction of the biological evolutionary framework as one which can reconcile the differences on the subject of preferences which have arised from mainstream economics and behavioural economics. Chapter four introduces the neurocognitive approach to the study of preferences, with the aim of investigating whether there are specific key regions in the brain which dominate the processes underlying the formulation of preferences. Chapter five contains the discussion, conclusion, and suggestions for future research on AI value alignment as a key security issue.

2.3 Scope and Limitations

Given that this is essentially a conceptual framework of analysis regarding an emerging area of security studies, one that concerns a hypothetical future instance which we are yet to solve, this research lacks empirical data and cannot currently extend to the realm of empirical experimentation. Furthermore, where it might be more beneficial to empirically test aspects of our preferences, doing so is beyond the scope of this paper. Thus, the case study method on preferences will rely on both primary and secondary sources. As this paper only extrapolates from the available literature on the topic, there are no ethical considerations.

Additionally, there is a tremendous amount of research on our preferences, coming from multiple disciplines. It would be beyond the scope of this paper to aggregate all of the findings to develop a formal general theoretical framework. To make sure that the paper does not lose its focus, however, the case study is bounded by the propositions mentioned earlier.

Chapter Three: Evolution of Preferences

Cambridge defines Preferences as "the fact of liking or wanting one thing more than another"

In this chapter I will attempt to pinpoint empirical basis under which our preferences emerge and have changed throughout time, by understanding their very nature and the nature of their origin, with the incorporation of a biological evolutionary framework. In this chapter I argue that, indeed, our preferences appear to be somewhat biologically and evolutionarily embedded within our genes, potentially providing a tractable empirical framework of analysis. Additionally, I argue that the specific notion of *evolutionary mismatch* may be key to understanding the divergence of our preferences, and enhance our ability to predict either further divergence in the future or the possibility of our preferences converging. By studying our preferences through the lens of biological evolution, we may then be able to extrapolate a more coherent and reliable theoretical framework that we can apply to current AI research on this subject.

In *Human Compatible: Artificial Intelligence and the Problem of Control*, Stuart Russell (2019) proposes a new way of thinking about AI which differs from the standard model of AI (which has been the center of focus in AI research since the inception of the field). Russell argues that the standard model of AI, which holds that the AI seeks to maximize *its* objectives, is deeply flawed. Instead, he reconfigures the problem of AI control by arguing that the new model of AI should seek to maximize *our* objectives (Russell, 2019; 11). The standard model of AI has been strictly dealing with fixed parameters, fixed reward functions, fixed environments and known goals. The standard model of AI works well with complete information; however, it is not sufficient in the sense that we can, through this model, develop a real-world applicable AGI that is aligned with our goals and values (Russell, 2019; 15-33). Considering that we, ourselves, are uncertain about

our goals, values and preferences, and that there perhaps could be no formal system of directly encoding said goals and values into an AI. Thus, the standard model of AI must change.

AI value alignment research was initially centered on constructing a general coherent framework about human values, in such a way as to potentially develop AI which can directly input and output our values without running the risk of misalignment which could lead to an existential threat. Building an AI with the slightest misconception of our values can launch the AI in an irreversible direction that is far-fetched from what our goals truly are and what we truly intend to achieve (Bostrom, 2014). The value alignment problem has proven to be exceptionally difficult to solve via direct value encoding. Thus, AI value alignment research is now shifting from the focus on configuring the very notion of our values and how to converge those values into universal human goals. Instead, research into value alignment is now focused on developing reliable methods of indirectly inferring our values from our preferences, by inferring our preferences from observing our behaviour. Human preferences give insight into how we come to value certain things throughout our lives.

Russell (2019; 174-186) proposes that a potentially more beneficial model could be one that indirectly infers our goals and values from observing our behaviours and by seeking our instructions, while remaining uncertain about our preferences. As opposed to the standard model of AI, this one would allow itself to be switched off, given that an uncertain AI cannot maximize an unknown and unfixed objective in a dynamic environment. Additionally, it would not create dangerous subgoals either, such as self-perseverance and unbounded resource acquisition; such subgoals (among others) would prevent us from controlling this AI (Bostrom, 2014; 109-114; Omohundro, 2018; 47-54). The AI's objective here is to realize its humans' preferences. If neither the human nor the AI can be absolutely certain about those preferences, then the AI refrains from

taking decisions towards maximization, and allows the human to turn it off, knowing that being turned off is well within the potential preferences of any human. The principles which constitute this model are the following (Russell, 2019; 176):

A. the machine's only objective is to maximize the realization of human preferences;
B. the machine is initially uncertain about what those preferences are; and the ultimate source of information about human preferences is human behavior.

The structure of the following section will investigate Russell's above-mentioned premises for the possibility of developing a beneficial AI through the lens of biological evolution. Neoclassical economics and behavioural economics have long debated the nature of human preferences and what constitutes rational choice, and have left the field divided upon itself with no general theoretical framework that truly captures our preferences. I propose the integration of biological evolution as a helpful tool provided by its empirical and tractable approach to bridge the gaps between neoclassical economics and behavioural economics and achieve reconciliation between the two. The incorporation of biological evolution can potentially illuminate formal pathways constituting the structure of human preferences. Additionally, this can shed light on the possibility of our preferences converging into something universal in the future on a collective scale rather than solely providing an understanding of individual cases. This approach might be helpful in the search for a coherent value alignment framework for AI, by tracing the beginnings of our preferences and the basic drives which had led to their current structure, and provide a more encompassing predictive power of our preferences on a larger scale.

In the past, the standard AI model of objective maximization, where the AI model knows its reward function and has complete knowledge of its environment and its parameters are fixed, is deemed obsolete since it is incapable of capturing the complexities and contextualization of our values as

ones which exist in the external world (Russell, 2019). AI research is now shifting into incorporating methods which seek maximization in unknown environments, with incomplete information and with an initially unknown reward function (Evans et. al., 2015; Woodworth et. al., 2018). This shift was primarily motivated by Inverse Reinforcement Learning (IRL) research which stemmed from psychology, where AI models now try to gradually learn the reward function by observing human behaviour, and later -from this- inferring human preferences. I argue that this shift, despite potentially being a step in the right direction towards solving the value alignment problem, remains predicated upon some flawed assumptions about our behaviours, which have originally stemmed from flawed assumptions about the nature and origin of human preferences.

The study of human preferences has been primarily theoretical, resting on assumptions regarding human nature and what constitutes rational choice for self-interested agents who seek to maximize their utilities as separate individuals. This was derived from neoclassical economists who proposed three primary axioms for our preferences, stating that a) we consistently choose the option which best maximizes our utility (albeit with some informational constraint); b) our preferences are stable; c) we constantly behave optimally (Weintraub, 2007). On the other hand, behavioural economists disagree on the statement of stable preferences by stating that they are, in fact, inconsistent, and that we consistently fail to choose options which best maximize our utilities (Kahneman and Tversky, 1979). However, both disciplines agree on the premise that human beings are naturally drawn towards maximization. In other words, maximization is an inherent feature of the human condition, yet we also systematically deviate from this. How true are those assumptions? Could they potentially be deeply flawed? A biological evolutionary framework has the capacity to shed light on these assumptions by studying the origins and the nature of our preferences and trace their change throughout time.

Effectively, humans now do not appear to show any consistent and universally encompassing goals and value structures. This is the pinnacle of the challenge regarding developing a value aligned AI: How can AI, AGI, or even superintelligence, be beneficially aligned to our values when we are not revealing a consistent goal structure, or a single global value or preference system? If there are indeed some formal underpinnings of human preferences that can be empirically tractable, then perhaps said formal underpinnings can shed light on the current shortcomings in IRL methods in AI research. If we are able to identify formal statements about our preferences, then this can potentially aid in the formulation of a general theoretical framework for AI value alignment. It sounds reasonable that, if we are to truly understand the nature of our preferences -in such a way as to be able to predict future preferences and how they may or may not converge- all to be able to infer our values, then we must go back to their very beginning in our evolutionary history.

Although biological evolution provides us with some beneficial statements about the nature of human behaviour and human preference, it cannot provide a comprehensive framework that can be utilized in forming individual predictions about single units in a population. However, while predicting the preferences of single individuals might be difficult given instability and stochasticity, I argue that with the aid of biological evolution, we can perhaps aggregate predictions of human preferences as a collective entity rather than separate and unique individuals. This entails becoming better at hypothesizing whether our preferences can converge in the future, or keep diverging further. While mainstream economics attempts to provide a mathematical explanation for a single ordering of preferences, evolutionary science may aid in not only the reconciliation of differences within mainstream economics and behavioural economists, but also potentially

provide us with information regarding collective preference ordering. This alludes to prospects of generalization via reconciling individual predictions with collective ones about our preferences.

3.1 Preferences and Maximization

In his book about artificial general intelligence, Tegmark (2017) touches on our intrinsic tendency towards maximization which, to him, appears to stem from the very basic laws of physics in nature:

> Out of all ways that nature could choose to do something, it prefers the optimal way, which typically boils down to minimizing or maximizing some quantity.. So if nature itself is trying to optimize something, then no wonder that goal-oriented behavior can emerge: it was hardwired in from the start, in the very laws of physics. (Ch. 7, para. 3)

In a comprehensive discussion on the difficulties of AI value alignment and the potential dangers of misalignment, this is the only indirect hint that Tegmark provides which seems to be directly connected to our evolutionary history. By virtue of inheritance, maximization represents a point of intersection between current research on AI inferring from human preferences and the emergence of our own goal-oriented behaviour as human beings, as we can infer from our own ancestral history.

Sewall Wright (1932; 357-360, as cited in Burnham et. al., 2015; 5), the inventor of adaptive landscape, argued that selection favors optimal fitness functions, i.e. enhancing survival. When the environment is in equilibrium (I.e. stable), individuals tend to cluster around a local maxima, where this local maxima represents the inclusive fitness function that enhances our survival. In contrast, when the environment is not in equilibrium, individuals do not appear to cluster at a

local maxima. Standard economics theory states that humans are in market equilibrium, thus we always opt for the optimal behaviour or decision considering our options and some knowledge constraints (humans are rational this way and otherwise considered 'irrational') and that we succeed at doing this. Biological evolution argues against this presupposed notion of market equilibrium as suggested by standard economics theory. According to biological evolution, the drastic change in our environment which began due to the first industrial revolution, happened too rapidly, in such a way that some of our adaptational functions had failed in catching up to the proceeding industrial advancements. Where humans were once successfully adapted to changes in the environment throughout natural selection processes (evolutionary match), we have also maladapted through the process of *evolutionary mismatch* (Burnham and Phelan, 2019a; 154; Wilson and Kirman, 2016). Standard economics theory does not take into consideration the fact that humans, from an evolutionary perspective, are -in fact- living in a novel environment in comparison to that of our ancestors. Thus, we are not in stable market equilibrium and our environment is novel.

Where environment equilibrium reflects organisms successfully maximizing their inclusive fitness function, evolutionary mismatch on the other hand, reflects how and why organisms fail to maximize their inclusive fitness function when they are not in equilibrium with their environment anymore. Burnham (2016a) believes that current economic strategies deal strictly in novel settings with novel problems that have only been recently invented. He argues that evolutionary mismatch can explain the irrationalities and inconsistencies underlying our preferences by saying that "Mainstream economics places no restrictions on preferences. Equilibrium evolutionary approaches suggest people will have preferences that create pleasure for fitness increasing behaviors, and create displeasure for fitness decreasing behaviors" (p. 199) Furthermore,

suggesting that “it might be natural to approach problems of evolutionary mismatch by first considering what behaviour would be favored in equilibrium, and then relaxing the equilibrium assumptions” (p. 205). In others words, by adjusting the stability parameters in an environment, one may come to observe the gradual change and divergence of our preferences more clearly. In turn, this can provide us with more accurate predictions of our preferences. For example, one may hypothesize that had the industrial explosions been stretched out over longer periods in time, our preferences would be more coherent and rational now, as they would be in equilibrium with the environment.

According to evolutionary research, humans at some far point in time were maximizing what mainstream economists considered as a quantifiable, measurable, and observable utility (Samuelson, 1948; 243–253). This utility simply was, during our ancestral times, survival and reproduction. Evolutionary scientists replace the term “utility maximization” with “inclusive-fitness function” to refer to survival and increased offspring reproduction as having been our inherent basic drives which gave rise to all of our current preferences (Burnham, 2016a). Survival and offspring production are more easily observable and measurable throughout our evolutionary history, as opposed to measuring subjective hedonistic values such as happiness or regret. However, as human societies started undergoing rapid industrial change, what were once observable and measurable utilities became unobservable, inconsistent, and unmeasurable. It is as if something had collectively tipped us off the scale of evolutionary adaptation and had changed our collective preferences, and in some aspects, it had even reversed some of them. Natural selection favors maximization; however, we are no longer maximizing some obvious utility, our preferences are no more accurately reflecting any specific “inclusive-fitness function” (Burnham, 2016a; 205; Hamilton, 1964; Trivers, 1971; Becker, 1976).

Harbi et. al., (2015) reveal through experimental evidence using both real and hypothetical situations that “even in very favorable conditions, people are not systematic efficiency maximizers. Individuals equally maximize efficiency and equality, and choose positional preferences to a lesser extent” (p. 83). This means that, in contrast to the assumption about self-interest and absolute utility (or absolute wealth), people actually reveal otherwise when put to the test in real-world contexts. The way we make choices, as if we are behaving in real situations, consistently differs from the way we behave in situations posed by mainstream economics, which are too ideal and mathematical. This begs the following question: if we are maximizers, but we are no longer maximizing offspring nor wealth, then what are maximizing exactly, and why do we sometimes fail at this? A point of reconciliation here, as Burnham and Phelan (2019a) explain, is that “natural selection does not produce behaviors. Rather, selection produces behavior-generating physiological mechanisms that are sensitive to environmental cues” (p. 146).

Take the example of our diet preferences. If the environment we live in is in equilibrium, then we are expected to eat foods that are healthy and energetic, ones which prolong our lifespan and produce more biologically competent future generations. Our current diet preferences, to an extent, reflect the opposite of this. The reason why we tend to prefer unhealthy food and unhealthy addictions (e.g. sugars and some types of fat) is because these types of foods were the optimal options available to our ancestors in an environment where food was scarce. Those past diet preferences had undergone many generations of selective pressure of maximization, thus becoming biologically embedded in the organism and later on is revealed in the notion of preference. In contrast, now we have an abundance of options regarding food, however, our preferences still match the preferences of our ancestors in this regard. Where our diet preferences could have evolved towards selecting for foods that increase our health, thus increasing our

reproduction and survival, we nonetheless appear to have failed -to an extent- in maximization here. The reason for this is because there have been zero generations of selective pressure to adjust our preferences toward fitness-increasing healthier diets that contain less sugars and harmful fat. Failure of maximization here is due to an evolutionary mismatch that causes our preferences to send the wrong signal. This mismatch was driven by the rise of agriculture and the proceeding industrial and technological explosions, which happened too fast for our preferences to evolve accordingly into true fitness-maximizing behaviours (Burnham, 2016a).

Take fertility as another example for maximization failure. Generally, the more resources given to an animal, the more offspring it tends to produce. In contrast, scarce resources lead to a decrease in producing offspring. With modernization rises an abundance of resources, however, fertility rates for US residents show the opposite of this. Greenwood and Seshadri (2002; 153-158) contrasted fertility rates of the US population from 1800 to 1940. In the 1800's, about ninety four percent resided in rural areas and women birthed seven children on average. In the latter, however, only forty three percent lived in rural areas and women birthed two children on average. The authors concluded that the driving force behind this sharp decline was technological progress. Similarly, Iron (1998; 195) further argues the same by explaining that the total fertility rates of three children or less would probably have gone extinct before this technological progress, considering the high death rates before the introduction of modern medicine. Indeed, natural selection favors more offspring, but the speed and level at which our environment has changed and morphed into something completely novel has produced an evolutionary mismatch on the rate of fertility, which is reflected now in our preferences.

3.2 Expected Utility, Trade Offs, and Discounting the Future

Daniel Kahneman (2011) provides a partial explanation as to why we sometimes make irrational decisions, or make decisions that are inconsistent with utility maximization. According to him, the way we make decisions is reliant on two primary systems of the mind (or the self): the remembering self and the experiencing self. The former is the one in charge of making deliberate decisions, while the latter evaluates memories only by their "peak" and "end" moments. In one of his numerous experiments, Kahneman showed how people's decisions can often be inconsistent with the principle of maximization. For example, in one experiment, the participants were asked to plunge their hands into cold water in two different regimes. In the first regime, the subject plunges her hands for 60 seconds in water at 14 degrees Celsius; in the second regime, she does so for 60 seconds in water at 14 degrees followed by 30 seconds at 15 degrees.

After doing both regimes, the participants were asked which one they would like to repeat (in other words, 'prefer'). The majority chose 60+30 over just 60 seconds in cold water. To the reader, this might not be your preference because spending more time in cold water is more uncomfortable than having spent less time. So, why did the majority of participants prefer prolonged discomfort as opposed to the shorter regime? Kahneman (2011) explains that the remembering self highlights the "peak" and "end" (or highest and lowest) points in memories, and pays special attention only to those points, effectively putting the remembering self (which appears to be irrational) in charge of making decisions. In both regimes, the peak discomfort was the same, but the ending was different for the 60+30 regime with water being 1 degree warmer. The rational decision should be to choose the regime with the lesser duration of comfort, but instead, participants chose the longer regime that had a milder discomforting ending to it. This,

among many other experiments, showed that the two selves are in conflict and are inconsistent, revealing a flaw in past assumptions about maximization and stability of preferences over time (pp. 378-385). Recent neurocognitive research supports these findings and is able to provide a biological basis for the underpinning mechanisms. Coolidge (2018; 82) refers to the function of the hippocampus as being crucial in emotional processing, revealing how humans tend to remember the things which incur an emotional valence and forget the things which do not incur this. This means that emotional valence plays a critical role in deciding whether or not we will remember something.

Schwartz (2004) refers to numerous examples in economics and psychological experiments to illustrate how and why our choice making mechanisms can be paradoxical or irrational. He distinguishes between “maximization” and “satisficing”. If you are a maximizer, then ”every option has the potential to snare you into endless tangles of anxiety, regret, and second-guessing” (p. 85). If you are a satisfier, then you can settle with the next best thing or what you consider to be ‘good enough’ when navigating complex decision problems. Schwartz (2004; 79-90) showed, by conducting numerous reliable scales that measure happiness, regret, and depression, that there is a consistent and statistically significant correlation between maximizers scoring low on all of those measures, and satisfiers scoring higher. The higher the maximization is, the higher the regret, depression, and unhappiness which the person experiences, and vice versa. In a sense, these results show conflict in the way we tend to make choices, where people who maximize tend to feel subjectively worse, as opposed to satisfiers. Aiming for the optimal decision does not always produce the intended positive subjective experience from that decision.

Furthermore, Schwartz (2004) explains how the abundance of choices that are available to people (whether readily or not) tend to negatively impact their overall subjective experience when trying to make the best possible decision, even though it is commonly established that more information begets optimal decisions about the decision problem, where those optimal decisions lead to higher utility. The reason for this dissonance is because human beings are simply incapable of computing highly complex and tedious problems in short or limited time frames in order to arrive at the best possible solutions which would maximize their subjective experience; the cost of brain computation can be quite high (William & Baumol, 1979). Experimental evolutionists, human behaviour ecologists, and evolutionary psychologists support this conclusion by contrasting variations of selective pressure in evolution to maladaptiveness in our current modern environments (Rayo and Becker, 2007; Clark et. al., 2008, as cited in Burnham, 2016a; 196, Tooby and Cosmides 1992, Barrett, 2012; Capra and Rubin, 2011).

One of the principles upon which standard economics theory predicates itself is Expected Utility Theory (EUT). According to EUT, when rational people make decisions under uncertainty and when facing risk, they tend to opt for the option which they believe will best maximize their future payoff. This entails a process for discounting the future. A simple way to illustrate this principle is that we tend to place a higher value on happiness today as opposed to happiness tomorrow; we are impatient. For example, when presented with the option of attaining a hundred pounds in forty days and a hundred and ten pounds on day forty-one, most people would choose the latter. However, on the fortieth day, people tend to reverse their preference and ask for the hundred pounds now as opposed to waiting one more day for a hundred and ten pounds. This preference reversal contradicts what standard economics theory presumes about the stability of our preferences. Behavioural economists were able to provide some basis for

explanation in our preference inconsistencies, yet without providing a fundamental and more general explanation for them.

There are two ways of discounting the future: hyperbolic discounting and exponential discounting. The former refers to the notion that people tend to wait less as the future nears the present by placing disproportionate value on the future relative to the present value. Exponential discounting, on the other hand, refers to a more consistent or formal model of placing value on the future relative to the present, and it can be represented in a stable mathematical fashion (Frederick et. al., 2002; 351-401). According to standard economics theory on discounting the future, a drug addict makes the rational trade off when she chooses pleasure today as opposed to a lifetime of misery (Becker and Murphy, 1988; as cited in Burnham, 2016a; 201). Much research has emerged which studies this in non-human animals. Not surprisingly, some types of non-human animals display consistent saving behaviours in terms of food foraging and storage for harsher times. Additionally, the way some of these non-human animals discount the future aligns with the expected utility maximization and future discounting as argued by proponents of mainstream economics. However, the way humans discount the future does not align with this. Natural selection agrees with the principle of discounting the future as it favors it. However, that is the case only if the environment is in equilibrium and if the objective of the maximization process is the inclusive fitness function, and not necessarily a quantifiable object such as wealth, which is what mainstream economists assume.

With the integration of experimental evolution, for example, we could be in a unique position to investigate the mechanism of future discounting by using empirical methods. In fact, this has already been done, albeit not with human subjects but with animals (Green et. al., 2010). It appears that animals, such as humans, do engage in behaviours relevant to discounting the future,

such as in the case of storing food for later consumption and delaying reproduction to increase future offspring lifespan (Vander, 1990; Rose, 1984). According to evolutionary mismatch, however, humans can save too little, too much, and sometimes showcase unstable preferences, effectively contrasting the principles of standard economics theory. Burnham and Phelan (2000b) explain the impatience in our preferences and our tendency for hyperbolic discounting by referring to our ancestral environment:

> Think back, once again, to that long period of our evolutionary history we spent as hunter-gatherers. We evolved in a world where wealth existed primarily in the form of food and could not be stored for very long — any surplus would rapidly spoil. So our brains were designed in an era when the best way to save was to consume. (p. 10)

Burnham and Phelan (2000b; 14-16) suggest that the reason as to why humans fail in maximizing their expected utility and display impatient preferences regarding discounting the future is because in our ancestral environment, having been harsh, uncertain, and providing only scarce resources, had forced earlier humans into developing preferences towards earlier consumption (where food was the currency) as opposed to saving for later times and risk spoiling food and not finding any more later. As our ancestors had nothing analogous to refrigerators or banks, the way to save for later was actually to consume now. By converting body fat into energy, this energy will later be utilized in future foraging to survive harsher days. When calculating risk in uncertain situations, humans tend to be rather 'bad' statisticians, where we tend to overestimate the odds of winning a lottery and underestimating the odds of other kinds of gambles (e.g. dying in a car accident is much more likely than in a plane crash). On the other hand, animals appear to be much better at picking the right strategy in either averting from risk or seeking it in order to obtain the optimal payoff, effectively appearing as better statisticians.

Take woodpeckers, for example. There was an experiment which was conducted on woodpeckers that were presented with artificial trees containing twenty-four holes in them. The purpose of this experiment is to know at which point would woodpeckers decide to desert a tree after encountering some number of holes. The idea is that a woodpecker would waste its time and energy going through all the holes in the tree to find food, but if a woodpecker leaves the tree too soon it may be missing out on food that could be in that tree. The experiment showed consistent results. Woodpeckers tend to abandon the tree after encountering about an average of 6.3 holes. This suggests that the reason as to why woodpeckers appear to be excellent at conducting their tradeoffs is because this knowledge is instinctive to them. They had faced the same problem for many thousands of years, thus, the solution to this problem had become hardwired in them. Animals tend to be rather good at doing these kinds of tradeoffs in their natural environment (Lima, 1984; 166-174). Evolutionary mismatch suggests that the reason why humans reveal suboptimal tradeoff strategies is due to us living in novel environments.

It might seem that this is a far stretched comparison, which can neglect various current contributing factors to the ways in which we exercise our preferences now; it certainly does not seek to provide an absolute causal biological explanation. However, the idea that our current preferences match the preferences of our ancestors can potentially be best explained by the idea that our ancestors' preferences became hardwired into their genes due to many generations of selective pressure to increase their fitness. Considering that we live in a relatively novel environment with relatively novel problems and situations, the inconsistencies in our current preferences may be better explained by referring to the fact that we have not undergone generations of selective pressure similar to those undergone by our ancestors, given that it takes a tremendous time for that to happen.

Burnham and Phelan (2000b) further elaborate on the power of our genes by comparing our reactions between novel dangerous objects and much older ones. For example, babies can play with a gun as if it was just another addition to their toys, yet immediately cower in fear if they are handed a plastic snake. Similarly, our closest relatives, the chimpanzees, display the same reaction immediately when presented with a snake, even though said chimpanzees have been born and grew up their entire life in a zoo and had never before seen a snake. This showcases the strength of our genes in determining some of our behaviours and reactions. Snakes have been not only our ancestor's enemy for many generations, but also an enemy to other non-human animals. Many generations of selective pressure perpetuated this instinctive reaction and can explain more of our current reactions and reflexes toward dangerous objects, in addition to the lack of said reflexes toward novel objects. They argue that our instincts towards investment and saving behaviours have not yet had time to evolve, given that money is a novel currency. Where in our ancestral environment, food would quickly lose its value over time because we did not have the technology to properly store it. Money, on the other hand, does not lose its value over time:

> Our ancestors knew nothing of financial instruments. Accordingly, we are no more likely to have instincts to make arcane financial decisions than we are to fear guns. Our instincts for saving for the future simply aren't wired for modern financial choices. Maybe in a thousand generations but certainly not this fiscal year. (p. 20)

Standard economics theory appears as if it is predicated upon how non-human animals discount the future, in this regard. However, the way that humans actually discount the future does not align in the same way as non-human animals do, meaning that the way we discount the future does not align with what the standard economics theory posits about stable preferences (Ainslie and Herrnstein 1981; Laibson 1997; Frederick et al. 2002, as cited in Burnham, 2016a; 202). Rubin and Paul (1979, 586-587) argue that cultural and psychological explanations of inconsistent

preferences (though in their paper they use the term “tastes”) can be useful in explaining only partial cultural and social aspects and partial individual elements. However, an evolutionary biological framework might be necessary to explain the larger sum of inconsistencies in all of our preferences on a collective scale.

3.3 Risk Preferences and Uncertainty

One of the key concepts in AI research is risk in uncertainty and the relevant notion of maximizing expected utility. AI researchers realize that we are not exactly rational, and that our risk preferences can sometimes be inconsistent and suboptimal. Evolutionary research may provide an explanation for this by studying how our preferences used to be rational or optimal in our ancestral environment, and how they later diverged from optimality into inconsistent risk preferences that do not maximize utility in our current time (Sharot, 2011). Rubin and Paul (1979) suggest biological basis underpinning our risk preferences. They trace our preferences for either risk-version or risk-seeking behaviours to our ancestral males, arguing how and why younger males tended to engage in risk-seeking behaviour, while older males displayed risk-aversion due to selective pressure that made such behaviours (in turn, preferences) genetically controlled over long periods of time. They argue that risk preferences (via unfair gambles) emerged to enable young males to join the breeding hierarchy, and if the risk behaviours they displayed for doing this persisted for sufficiently long periods of time, then those behaviours become “programmed in our genes” (p. 590) and later revealed in the form of preferences.

Furthermore, our ancestral behaviours towards risk-aversion relates to future discounting, as mentioned earlier. Rubin and Paul (1979) believe that even though increasing offspring reproduction better ensures the survival and continuation of one's genes, our ancestral males of older age, unlike the younger ones, were risk averse. They averted from risk behaviours associated with acquiring another mate and producing more offspring given that the older male may not survive long enough to father the proceeding offspring, in addition to the mortality rate implications at that time. Thus, the older male values his existing family more than a potential future one, since there would be no positive returns on a future family that may not survive without a father to sire them. According to mainstream economics, when one gambles with risk, they either immediately acquire a higher utility (wealth) or immediately lose it. This notion is contrary to how natural selection works, because the payoff of inclusive-fitness function is not immediate, it takes much longer before it becomes evident. Thus, risk preferences which immediately double wealth do not necessarily double fitness on the long run (p. 593-594).

According to this line of reasoning, gambling addiction is rational for mainstream economists, as the gambler successfully achieves optimal pleasure now over a lifetime of misery. However, it is not rational according to natural selection, since the consequences and effects of gambling can induce effects that are detrimental to the welfare of the individual, thus negatively affecting her fitness in the long haul. The relationship between wealth and fitness may have been linear in our ancestral environments, however, the way our preferences have become maladapted due to evolutionary mismatch has now rendered this relationship nonlinear.

There have been studies which assessed the degree to which our genetics can determine our preferences. For example, Cesarini et. al. (2009) conducted twin case studies to assess the extent to which genes determine twins' preferences for risk, and to empirically explore the source of

preferences heterogeneity. Regarding allocating assets, they found that there is a correlation between zygotic twins (I.e. identical), where said twins revealed the same preferences. Furthermore, Cesarini et. al. (2009) argue that any attempt to understand heterogeneity in preferences is bound to be incomplete without taking into consideration that our preferences could be under some genetic control (pp. 809-842). If our preferences were more or less determined by genetics, then this could be the case due to the preference behaviours of our ancestors that had become embroidered in their very genetic makeup due to having passed many generations of selective pressure on those preferences. In other words, their preferences were the result of an evolutionary process (Levy, 2010).

Additionally, Hintze et. al. (2015) hypothesized that the reason as to why humans evolved risk-sensitive (e.g. risk-averse) adaptive behaviours was due to having lived in small groups (about 150 individuals), where larger groups produced no specific risk-preference. By using a genetic algorithm, they conducted agent-based simulation to study the effect of group size on the evolution of risk-sensitive strategies, by evolving strategies that adapt to playing the equivalent mean payoff gamble. This gamble entails that the reward payoff is the same when either choosing the risky option or the safe one, where the safe option gives a fixed payoff with one hundred percent certainty, and the risky one awards a higher payoff but only half of the time. The study concluded that preferences for risk sensitive strategies depends on the group size, but not the population size. Groups of smaller size effectively evolved risk sensitive preferences, while groups of larger size evolved no such preferences. Even when the gamble is iterated for a number of times and when there were migrations between groups, smaller groups still evolved risk sensitive strategies. Their study also produced novel results regarding the introduction of mutations, which

has not been done in earlier studies. They revealed that, even when incorporating mutations, risk-sensitive strategies in small groups still evolved (p. 5).

On a final note, I will briefly discuss the notion of biases and how it fits into the evolutionary explanation of our preferences. Capuchin monkeys and humans appear to show strikingly similar dispositions in choice action, this is believed to be due to the fact that humans and capuchins share a common ancestor from thirty-five million years ago, as this common ancestor exhibited similar choice actions. This means that our choice strategies are quite old in an evolutionary sense, rendering them somewhat outdated. This might explain the notion of biases pervasiveness in our current environment, as such biases have been selected for over the course of millions of years, thus enabling said biases to become genetically controlled in some aspects. This, in conjunction with the data presented in this chapter, may be quite helpful in informing policy makers to utilize the evolutionary framework into their efforts at understanding, modifying, and predicting human behaviour, specifically with regards to our preferences and decision-making processes (Lakshminarayanan and Santos, 2011; 75-89).

Thaler and Sunstein (2008) suggest using “nudges” as means of working around our biases and predispositions, as opposed to attempting to directly influence and change them. The Nudge method proposes that choice and preferences can be influenced by presenting indirect suggestions and positive reinforcement. Additionally, Tversky and Kahneman (1981; 453-458) suggest using “framing effects” as a potentially beneficial strategy to avoid the pitfalls of our predispositions, inconsistencies, and biases. The framing effect explains how our choices and preferences can be influenced by the specific ways in which the available options are linguistically phrased, while also changing the settings context of the situation or problem.

This chapter investigates our preferences through the lens of biological evolution, and specifically evolutionary mismatch. It is believed that many of our inconsistent or irrational preferences have persisted to this day due to the inability of our genetic composition to catch up to the rapid environmental changes that have occurred over the course of our history. The first major change happened about ten thousand years ago, with the introduction of agriculture and farming. Several major changes followed, from the first industrial revolution to the current technological explosion. The idea is that our preferences were once consistent, convergent and rational, when we lived in equilibrium with the environment. However, our preferences have diverged into inconsistency and irrationality, as we no longer live in equilibrium with the environment. The evolutionary framework can be quite beneficial when seeking to understand the current divergence of our collective preferences as a species, in addition to our individual inconsistencies, by providing a kind of a causal link that is tractable through our evolutionary history.

At the very least, evolutionary research may be able to provide us with the beginning of a thread that may later on lead to developing an ultimate causation framework, or something close to that effect. At the very best, evolutionary research can help AI researchers in providing better and more inclusive predictions of our collective preferences, and how they may or may not converge to a coherent goal structure in the future.

Additionally, when looking at our preferences through the lens of evolutionary mismatch on one hand, and looking back at the value alignment problem on the other, one may notice an interesting observation. While evolutionary mismatch argues that rapid technological growth has caused a tremendous divergence in our collective preferences and values, AI becoming misaligned to our values due to rapid self-recursive improvement on itself sounds analogous to the former statement. In other words, the ways in which AI can become misaligned to our values,

is similar to the ways in which we have become misaligned to what could have been our collective coherent values and goals now, had rapid technological growth not occurred so drastically and so fast, in such a way that it had prevented us from catching up to it and retaining beneficial collective preferences, values, and goals.

One cannot fully capture the meaning of human preferences and be able to predict those preferences by relying solely on what the economics and behavioural literature has provided for us. Incorporating a biological evolutionary framework can aid our understanding of the true nature of preferences, how they originated, and how they may or may not converge into collective coherency in the future. The evolutionary framework can provide us with a better predictive power regarding our collective values, and this may, then, provide us with a better theoretical AI value alignment framework that takes into consideration all aspects relevant to the formulation and adaptation of our preferences. Experimental evolution can be particularly useful to empirically study the axioms in mainstream economics. For example, researchers can design adaptive landscapes that can empirically trace those axioms, such as the origins of our tendency towards maximization (Burnham et. al., 2015; 6).

Chapter Four: The Neurocognitive Science of Preferences

Studying the nature of preferences requires the investigation of potential neurocognitive markers which underpin their formulation. This is because said preferences cannot overcome the ways in which our various brain regions function. Neurobiological research has aided our understanding of preferences, to an extent, by studying the independent functions of the subsystems of our brains, in addition to the interconnectedness of those subsystems. While researchers in the past have argued for the existence of specific regions within the brain which may exclusively affect independent behaviours and psychological phenomena, recent research produced numerous findings that are contrary to this (Wilson and Kirman, 2016). For example, research on brain modularity (I.e. the study of specialized regions within the brain which give rise to specific cognitive functions) proposes that there is no independent region within the brain which takes full control of any specific cognitive function (Arnemann et. al., 2015; Sporns and Betzel, 2016; Kelkar and Medaglia, 2018). Instead, each cognitive function is subject to the simultaneous interaction of multiple regions. From this, it is suggested that preferences do not casually arise from a specific brain module, rather that preferences represent the very notion of connectivity itself between different brain modules. This poses serious questions in terms of understanding how to develop an AI that infers our values from our preferences. In other words, it may be much more challenging (if not impossible) to code the notion of connectivity itself in an algorithmic way, rather than coding a specific independent function of the brain. If our preferences represent the strength or weakness of connectivity between different brain

modules, then how can we compute this in an algorithm in order to replicate the connectivity process?

Furthermore, there are specific cognitive functions which may greatly change our understanding of preferences, alluding to preferences being much more amenable and subject to manipulation than we had initially thought. For example, much research has emerged which studies the nature of our memories and how they affect our present and future preference formulation, suggesting that our preferences are far from stable and do not mirror truthfulness from the reality in which they emerged (Johansson et.al., 2008; Mlodinow, 2012; Tekin and Roediger, 2017). This occurs due to memory confabulation, partial false reconstruction of events, and biases. This idea may render the concept of preferences far from any concept which we can mathematically understand and thus algorithmically replicate. This chapter aims to illuminate three key neurocognitive aspects deemed necessary to understanding the formulation and adaptation of our preferences that stem from neurocognitive research. In this chapter, I achieve this by focusing on the interplay between preferences and memories and intentions, memories and choice sets, and brain modularity.

4.1 Preferences: AI Modelling, Where We Are Now

Yudkowsky (2004) proposed the concept of Coherent Extrapolated Volition (CEV), which essentially discusses the possibility of creating a friendly AI that would become aligned with human values as they converge for common universal goals. To understand CEV, one would ask oneself: What would our goals be and what would we do differently if we were able to think better and faster? How would our goals differ if people had grown up closer together or farther away from each other? This proposition hints at the idea of an AI being able to predict an idealized version of us, based not on what we currently have, but what we could have had if

things were different. Yudkowsky (2004) almost immediately dismissed this proposition after publishing it, arguing that our values and goals may not even converge, coupled with the idea that an idealized collective image of our values is riddled with many fundamental questions, and that the very notion of writing this kind of complex AI program could be lightyears away from us. It is true that inferring values and goals which unite the collective preference ordering of humanity has proven to be an exceptionally difficult task for a very long time. However, I argue that a new model of AI which seeks to infer our goals and values from our preferences, even when it is uncertain about our preferences and does not maximize, as suggested by Russel's propositions, is just as difficult as the value alignment problem itself. You may call this *The Preference Alignment Problem*.

Parallel to what Yudkowsky (2004) proposed, human preferences, such as human values, may not even converge to universal preference orderings. Additionally, designing a friendly AI program that analyzes human preferences will be just as difficult, if not more than designing one which can predict and align to our values. The cognitive structures which underlie preferences appear to be even more complicated and ambiguous than those constituting goal-oriented behaviour and the formulation and adaptation of values (Symmonds and Dolan, 2011). One may have an overarching goal of becoming a famous actor, for example, but there will be a tremendous amount of both hidden (unconscious) and obvious preferences and sub preferences which will, in an inconsistent and irrational manner, direct the behaviours and decisions of this individual. To complicate this further, some aspects of those preferences may be completely inaccessible through conscious deliberate thought.

The idea of creating artificial agents which seek to model our behaviours and preferences has become a central area of research in the Value Alignment problem. The idea is that if an AI can accurately reason about us, then it can reliably infer our goals and preferences (Fallenstein and Soares, 2018; Yudkowsky and Marcello, 2013). There has been much ongoing research on various AI agent modelling techniques or strategies which reason about other AI agents, such as plan recognition, where an agent identifies another agent's goals via hierarchical action descriptions; recursive reasoning, where an agent tries to predict another agent's actions by its beliefs and the beliefs it ascribes to other agents; group modelling, where an agent creates joint predictions about the decisions or behaviours of an entire group or subgroups of agents (Albrecht and Stone, 2018: 66-95). Those are only a few modelling techniques among many others. Any one of these techniques can be effective within its own fixed domain; however, there is as of yet no dominant modelling technique which can be applicable across the board, and certainly not one which can transcend into the complexity of the real external world.

There are currently a number of problems regarding those modelling techniques, such as computational complexity, where the computational resources are insufficient for solving problems of long and complex decision trees and recursion. Additionally, there are observational constraints, where the agent cannot reliably reason and infer about other agents when presented with only partial information about the agent and the environment (Albrecht and Stone, 2018). The crucial problem that is shared within all of these modelling techniques, with regards to value alignment to humans, is that none of these models can be reliably applicable in real world scenarios. The reason for this is mainly because none of these models are trained to hypothesize about real agents (humans) in highly complex and context sensitive situations (Tegmark, 2017b:

355). Thus, it is hypothesized that if these problems are solved, then agent modelling can be transferred reliably into real-world situations.

The above-mentioned shortcomings of current work in AI modelling refer to what is called "Kolmogorov Complexity" and "Combinatorial Explosion" (Bostrom, 2014: 10-11). The former, also known as Algorithmic Complexity, refers to the idea that if it takes too long to describe a random object then this would effectively lead to a recursive impossibility. A human being can be perceived as recursively impossible to an AI that seeks to understand this human being and infer her goals and values from her decisions and behaviours. A simple object, on the other hand, is as such only if it can be described with just a few words. A complex object would need a much longer description, or potentially appear to be impossible to describe (Wallace and Dowe, 1999). The reason why humans are considered highly complex is because — to the best of our knowledge— our brains do not appear to be fully computational; if they had been so, then we could more easily replicate or simulate them perfectly and represent them in a pure formal style. In a sense, saying that our brains are fully computational would mean that our brains are not 'random' and are 'simple'. However, extensive research in psychology, neuroscience and cognitive science show otherwise (Chalmers, 2012: 214-248; Naselaris *et al*. 2018: 365-367; Kriegeskorte and Douglas 2018: 1148-1160).

Many researchers believe that our brains are somewhat computational (in that some mental processes can be simulated and their neurological underpinnings understood), however, there are still many questions regarding higher cognitive processing such as free will, consciousness, and intentionality. There has been extensive research on consciousness coming from multiple disciplines, some hinting at some underlying cognitive properties and their behavioural relations, but none that reveal an encompassing and clear systematic process(es) as to how consciousness

actually works and what it is (Searle, 1980; Leisman and Koch 2009; Chalmers, 1995: 212). Having said as much, if an AI or AGI (even if it is sufficiently intelligent) tries to model or understand our preferences so that it can infer our goals and values from them, it could still nonetheless face the obstacle of a Kolmogorov complexity, which would prevent it from fully capturing our intentions and values from the information it retains about our preferences.

Combinatorial explosion is similar to Kolmogorov complexity. A problem or an object can be described as combinatorial if it rapidly becomes more complex with more informational input (Gass and Harris, 2001). This suggests that an AI can face a combinatorial explosion with more and more input about us and our preferences, in its attempt to grasp our intentions and values; which we, ourselves, are uncertain and sometimes even irrational about. There has been a lot of research in economics specifically, and psychology, which show that people tend to behave irrationally, or in other words, we tend to go against what is logically and rationally the best course of action (Tversky and Kahneman, 1992: 297-323; Ellsberg, 1961: 643-669; Bortolotti, 2004: 359-375). This is true about us, not only when navigating popular moral dilemmas, such as the Trolley problem, or decision-making ones such as the prisoner's dilemma. People tend to sometimes choose irrational courses of action even when the immediate and long-term payoffs are clear and the given situations are simpler, such as in the case of a student who knows she needs to pass her final exam but nonetheless chooses to procrastinate studying for it until the day before the exam is due.

Neurocognitive research has long been exploring behavioural modelling, suggesting that the brain contains a mirroring system, which fires the same neurons both when the human or animal is performing a behaviour, and is watching the very behaviour being performed by another human or animal (Keysers and Gazzola, 2014). It is argued that this neurological phenomenon is

particularly relevant for understanding other people's actions and learning those actions and intentions via imitation (Heyes, 2010: 575-583). This represents a point where work on AI and neurocognitive science intersects. If an AI can learn our preferences by observing our actions, in the same or a similar manner as how we and other primates do, then there is hope for developing a beneficial AI that poses no existential risk to us as it exponentially becomes smarter and more resourceful.

There has been extensive research on behavioural modelling and inferring intention from observing behaviours, stemming from both cognitive psychologists and social neuroscientists (Premack, 2010: 22-32; Cubitt and Sugden, 1998: 761-771). For example, the theory of mind, proposed by David Premack (2010), argues that humans can ascribe thoughts and intentions to others, to predict and explain their actions, and to postulate their intentions. According to this, we are then able to interpret and predict the behaviour of others, in addition to better understanding our own behaviours. If there is an underlying formal mechanism by which humans infer about each other by observing behaviours, then perhaps the same formal mechanism can be computed for an AI and make it behave more reliably in the same manner that we do, accurately inferring our intentions and preferences from our behaviours. However, inferring our goals through imitation of our behaviour has proven to be a challenging problem in the field of AI and robotics, where this leads to problems such as the aforementioned ones like combinatorial explosion and Kolmogorov complexity (Yudkowsky, 2004).

If we want to understand the formulation of human preferences in order for us to gain a better understanding of how AI should model those preferences, then it is best to look at the literature of our memories. The functions of our memories have been the subject of vast research for the past decades. Luckily now, with new technologies such as functional magnetic resonance (fMRI),

we are able to show consistent, reliable, and more empirical results regarding how our memories shape our preferences.

4.2 Preferences: Memories and Intentions

As opposed to complete deliberation and optimal choice behaviour, Chaiken and Trope (1999) proposed dual-process models of persuasion, which showcase the fact that contextual cues from our environment tend to influence some of our behaviours and the choices we make by triggering automatic responses in our behaviours. Emerging neural evidence now supports this proposition; the reason why we tend to deviate from rationality and optimality is because we are susceptible to the subliminal cues produced in our dynamic environments, of which we cannot immediately (or ever) completely and rationally comprehend in such a way as to intentionally produce optimality in behaviour and choice (Thaler and Sunstein, 2008). In other words, while some aspects regarding the ways in which we engage in decision-making and reveal our preferences are subject to rational and conscious deliberation of thought and action, other aspects of this same process are mediated instead through automatic responses, which are triggered when we absorb subliminal cues from our environments. These automatic responses escape our conscious deliberation and thought. Effectively, this means that our cognitive architecture prevents us from reaching optimal choices and developing rational preferences as we are incapable of acquiring complete information from our surroundings in full awareness. Thus, we are in a sense, always in the dark when making choices, and when attempting to rationalize the origin of our preferences and why we have developed those preferences, alluding to the fact that we are always in a state of uncertainty.

An interesting change in preferences emerges when we try to bring their origin to light, or to conscious thought. Various researchers, coming from marketing and consumer psychology, studied why the preferences of people tend to change when they are specifically asked to try and clarify the intentions underlying them. This is referred to as “the mere-measurement effect” (Morwitz & Fitzsimons, 2004; 64-74). It seems that this change tends to come about when people try to engage their automatic processes with their conscious thought, thus arriving at different conclusions regarding their preference intentions. This, in turn, causes them to change their mind about the nature of their preferences as they become more aware of the factors that led to formulating those preferences. The change in preference perception appears to happen due to simply being asked about our intentions. This alludes to the notion of our preferences being highly amenable to change and inconsistencies upon merely requesting some level of introspection (Dennet 1991; 88-89).

Johansson et. al. (2008; 142-155) studied this by conducting an experiment where the participants were presented with pictures of two female faces and then asked which face they thought was more attractive. After choosing, the participants were immediately asked to explain their choice. Later on, the experimenters would quickly change from the photos which were chosen into the photos which were not chosen, and present the latter to the participants, again asking them why they preferred this photo, even though it was not the one they had initially chosen and it looked quite different than the one they did. Many of the participants did not notice the change, and proceeded with justifying their preference for the photos they had not actually chosen at the beginning of the experiment. In fact, seventy five percent of the participants were blind to the change. The interesting thing was that, after the experiment had ended, the participants were asked whether they thought they were capable of noticing a change in a

hypothetical experiment where they would be presented with a photo that they chose, but later on would justify their choice when looking at a completely different picture. Eighty four percent of participants said that they felt confident that they would notice the change, in other words, they wouldn't be blind to it. When the results of the experiment were shared with the participants, they were surprised and some even expressed disbelief. The findings here support the idea of memory confabulation (I.e. constructing completely false recollections). Not only this though, but also support the notion of people tending to resist their own false recollection later on when presented with the truth. This means that our future preferences are potentially greatly dependent on our memories, and those memories could be far from true when compared with the actual events that took place. This phenomenon is termed "the introspection illusion".

John Searle (1984a; 1989b) argues against the idea of AI attaining a state of consciousness and showcasing intentionality such as we do. He says that, while AI may exhibit some signs alluding to the possibility that it thinks in ways similar to the ways in which we think through language, it nonetheless will not and cannot morph into the same kind of conscious entity that we are. Things humans exhibit, such as intentionality, cannot be replicated in a purely mechanical system. The study I mentioned above alludes to what Searle argues against. If AI is a threat to humanity, it will be so because it misunderstands our intentions, thus it cannot fully capture what we truly want and do not want. According to the study mentioned above, and numerous other studies on our preferences and their amenability, perhaps we need to revisit the concept of "intention" and provide a better and more empirical clarification for it. Otherwise, we risk developing an AI which may never be able to capture our intentions, given that we, ourselves, are uncertain and potentially highly ambiguous about. The current problems we are facing in building this kind of

AI revolve around robustness and a deep ambiguity regarding clear identification of concepts and problems about our preferences and behaviours.

4.3 Preferences: Memories and Choice Sets

Kahneman (2011) showed that the way we make decisions is dependent upon the relationship between two systems in our brains, or two selves: the remembering self and the experiencing self. Similar to this notion, argues Leonard Mlodinow (2012), that our behaviours are dictated not only by the conscious, but potentially to a great extent by the unconscious as well (including subliminal environmental cues). There have been countless psychological and neurocognitive studies that have revealed (and to an extent also measured) the underlying relationships between our behaviours and our corresponding brain processes, of which we are potentially highly unconscious of (Loftus, 2005: 361-66; Cesi et. al., 1994: 388-407; Freedman et. al., 2001: 312-16). Regarding forgetfulness, memory retention and recall, for example, research on the subject consistently shows that our brains are greatly amenable to an active process of 'gap filling' in our memories, or rather, memory reconstruction (Sporer, 2008: 737-57). As Mlodinow (2012) argues, much of these gap fillings are dictated by what goes on in our unconscious. The unconscious (or as referred to now as the new unconscious) refers to the various cognitive processes which are not accessible to the conscious mind due to the complex architecture of the brain. With the introduction of new technologies, those unconscious processes can now be revealed by studying the activation of the various regions in the brain, in addition to the interconnectedness of those regions and how they are activated simultaneously.

For example, Bartlett (1932; 68) shows how the recollection process of our memory actively changes the constructs of the memory itself and creates deviations of it throughout time; Wade et. al. (2002; 597-602) reveal how we are capable of manufacturing completely false memories by showing false photographs to individuals and then asking them to remember the event from those pictures; Nickerson and Adams (1979; 287-307) show that there can be great disparity between what we perceive, register, and then ultimately remember. Each of those studies have been replicated throughout the decades, some adding and others subtracting variables, yet nevertheless all showing surprising consistency in the results obtained. Modern research supports those earlier findings with technologies such as fMRI (Tekin and Roediger, 2017; Abe et. al., 2008: 2811-2819; Finley et. al., 2017: 154-168).

The way we construct our preferences is thought to be based on prior knowledge, experiences and expectations, coupled with our natural tendency for maximization or optimality. Given that we are in a continuous and active process of preference structuring, our preferences may be largely based on gap fillings, confabulation, and prolonged distortion of experiences and past events, meaning that our preferences are highly amenable to manipulation and could be potentially greatly inaccurate. In other words, the process which formulates our preferences cannot escape the processes of memory formulation; they are directly interdependent. To illustrate this, one can ask: What would my preferences be if I had a perfect recall of all past interactions and all of the knowledge gained? We may not know exactly how our preferences would diverge from what we currently value, but we can be rest assured that they could, nevertheless, greatly shift from what we presently have and desire if we had perfect recall. The shortcomings of our memories, which are inaccessible to our conscious mind, can make the

process of modelling human preferences, thus human decision making, very difficult and tricky, if not impossible.

Numerous studies on memory reconstruction have shown that people tend to be reluctant, or even resistant to accepting their own first account of a specific memory, in favor of later recounts of the experience or the event (Neisser and Fivush, 1994). The reason for this is because, in the process of repeated memory retrieval, our brains tend to omit details which we deem as irrelevant, inconsistent or 'quirky'. Then, our brains fill out the gaps left in place of those details with ones that appear more logical to us, in order for the memory to make more sense. Many studies on this subject showed that we tend to be more incorrect or false than we initially assume about how we perceive past events, with some experiments even showing that we are easily prone and more than capable of constructing entirely false experiences and memories (Roediger, 1974; Roediger and Neely 1982; Roediger and Gallo 2004) . Much of what happens in courtrooms, for example, regarding eyewitness testimonies and offender identification, alludes exactly to this (Wells and Olsen, 2003; 91-277).

If some or all of our preferences have been predicated upon prolonged memory distortion, alterations and false reconstruction, then, how reliable could our preferences be to a version of us which did not suffer from these shortcomings? An AI that studies our preferences would not suffer such as we do when it comes to storing information, deciphering and later retrieving it. Logically then, should the AI construct a better and more accurate mental model of our preferences? More importantly, would we even feel inclined towards accepting the AIs 'unbiased' suggestions of what could have been our preferences and goals, or would we resist those suggestions, in the same manner we resist them when they arise from our own later

recounts? Those are all crucial questions that need to be addressed, as they refer to more points of contact between this kind of AI decision making model and that of ours.

Coolidge (2020) argues that non-associate (I.e. associating stimuli without reinforcement) learning preceded associative learning (I.e. associating stimuli with reinforcement), by going all the way back to the origin of life itself on earth and observing the patterns of survival as having first originated through non-associative learning, and that associative learning had emerged afterwards. He says that the earliest molecular forms have survived due to two primary abilities: a) "to ignore stimuli not in their metabolic interest (I.e. habituation); b) to approach stimuli in their metabolic interest (I.e. sensitization)" (p. 33). In other words, survival was driven primarily by the ability of an organism to ignore certain things in its environment, to better enhance its capability at pursuing the things which would increase its survival, without the need for reinforcement to strengthen the associations between various stimuli. According to this, non-associative learning was the first fundamental driving force that had later led to evolving the ability for associative learning and had evolved our memory functions.

An AI that models our human preferences might need to engage in a similar process. AI must be able to perform efficient categorization of our preferences. Which of our behaviours and decisions fall within the domain of preferences for academic achievement, for example? Is it even possible to categorize preferences? People tend to have overarching preferences (such as long-term social status) and sub preferences (e.g. acquiring social status through political engagement) as opposed to academic achievement or political power. Some preferences may be stationary (e.g. I always opt for dark chocolate) and others entirely contingent on the exact moment of time and place (e.g. right now I would enjoy a vanilla ice cream more than a strawberry one). The idea is that the concept of preferences is vastly encompassing virtually all

human choice activities and decision-making processes; thus, making the notion of AI reliably capturing those preferences very challenging, not to mention the challenge of capturing those preferences without further inadvertently enhancing their pervasive underlying biases.

AI value alignment research has now shifted to using Reinforcement Learning, which was initially a psychological technique devised by the father of behavioural psychology. Skinner (1953) formulated the theory of reinforcement learning, proposing that the offering of rewards as means of conditioning a desired behaviour will increase this behaviour. On the other hand, subjecting the individual to a form of punishment will lead to a decrease in her undesired behaviour. AI researchers adopted this technique and implemented it in their current work. The idea is to create an AI where its actions are mediated in the same manner as our behaviours as mediated, by creating a reward function for the AIs desired actions, and a punishment function to reduce its undesired actions. Gradually and with practice, the AI will become better at maximizing objectives, as it seeks to obtain its reward.

Now, AI researchers, seeking to navigate the difficulties of uncertainty and inconsistency in modelling our own behaviours and preferences, have resorted to another dimension of the same technique: Inverse Reinforcement Learning (IRL). This technique, as opposed to the first one, entails that the AI does not begin its problem solving with a known reward function and can only partially access this reward gradually by trial and error, as it learns more and more about the problem which is given to it. Thus, instead of maximizing objectives to obtain the reward function (where the AI knows the reward), now the AI needs to learn the reward (it is initially obscure to this AI) by inferring it through observing behaviour. In other words, learn the reward given the behaviour (Russell, 2019).

There is an important shortcoming to point to here. While AI researchers seek to develop an AI that will, over time, infer our values by observing our behaviours and inferring from our preferences by using IRL, they do not mention another important finding that relates to Reinforcement Learning: Post-Reinforcement Pause (PRP). PRP states that, even when the reward is highly desirable, if the amount and time of effort that is put towards attaining it is significantly large, then even when attaining the reward, its initial subjective value has been already lost in the process (Griffiths and Thompson, 2017). Thus, the effort put into trying to reach the reward becomes valued more than the reward itself, which produces negative feelings in the individual at the expense of the computational cost required by the process (Mlodinow; 2012; 48). PRP is strict to the human condition, given that navigating problems incurs a computational cost that hinders choice optimality. That being said, PRP would not apply to AI, as the computational cost for AI will be either very minimal or non-existent. The problem here is that, as humans undergo more and more PRP throughout the various choice sets they encounter in their lives, the way they value future choices will become, to an extent, highly dependent on PRP memories, thus their values will change accordingly.

Louie and Glimcher (2011) aggregated findings which support the notion of PRP, with studies that aim at exploring the size of choice sets on the actual choice behaviour and the corresponding neurological activation. It appears that, the larger the set size of available choices, the more the individual is prone to ignoring or neglecting the action of navigating the entirety of the choice set. In contrast, the smaller the set size, the more efficient the decision-making process can be. This contrasts earlier speculation that more information necessarily begets better choices about a given situation or problem. We tend to opt out of lengthy and complex information in favor of simpler and lesser information to reduce the computational cost that our brain would undergo

in order for it to conclude with the optimal choice, which affects the ways in which we will be valuing the same or even different choice sets in the future, effectively shaping our preferences. This echoes earlier findings about our failure of maximization (Coolidge 2020).

Mlodinow (2012) says "we should maintain our appreciation of the fact that, if our mind's natural view of the world is skewed, it is skewed for a reason" (p. 199). When our brain omits certain details from our memories, it does so for good reasons. According to research in evolutionary psychology and neurosciences, the reason why we do not have perfect recall of events is to enhance our chances of survival. For example, if you are in the wilderness and you come across a tiger, you do not pause and try to examine the tiger: '*is it truly a tiger because it has this number of stripes? What is the statistical evidence or probability of encountering such an animal in this exact place and time? Is it a tiger because it is roaring at me?*' No. You simply run. You can contemplate the event later when you are safe, but for the time being, your brain does the job for you and concludes that there are good chances that this big and 'scary' (and potentially hungry) animal will try to make you into its meal. Our brains have evolved in such a way as to become better categorizers; you do not need to be conscious of the entire history of evidence supporting the dangers of tigers when they encounter humans and other vulnerable prey in the wilderness (Freedman et. al., 2001). To process information faster and more efficiently, you only need the 'gist', so that then you are able to quickly conclude that '*tiger is bad, must run now*'.

There have been many findings which emerged in support of what Mlodinow (2012) alluded to. Daniel Schacter (2012) perceives what initially appears as cognitive myopia and shortcomings from our memories and overall cognitive functioning not as evidence of any flaws, rather that these 'shortcomings' are actually evidence of efficient cognitive functioning that enables us to take in relevant new information and disregarded irrelevant information that may hinder our

decision-making processes. Damasio et. al. (2008) and Damasio (1994) suggested the somatic marker hypothesis, which states that our emotions are in fact, highly relevant to the ways in which we make decisions and formulate our preferences. The somatic marker hypothesis, echoing emotional valence (Bradley et. al. 1997), relates to the idea that when we encounter competing options, we tend to attach emotional charges to those options, and, in turn, this helps us better navigate different possibilities and make decisions more efficiently in the future. Similarly, and echoing Schwarts' (2004) comparison of maximizers and satisficers, Simon (1955) proposes the satisficing model, referring to reservation utility, meaning that our brains optimal strategy is in fact, to reach the first level of utility which satisfies us, rather than navigating all the options in hope of finding the optimal choice. Doing this would incur a computational cost on us which would reduce our subjective experience of the optimal choice, had it been obtained. Sharot (2011) echoes the above findings:

> Emotion enhances the vividness of memories, which are the ingredients of future projections, resulting in simulations that are more exciting than real life and consequently in over-estimation of our reactions. Motivation biases imagination such that simulations of positive events are especially compelling, which is partially why humans tend to be unrealistically optimistic. (p. 68)

It is important to provide a more comprehensive framework on the extent to which our perception of computational cost can actually affect and change our future preferences. Going back to the question I posed before, "what would you prefer if you had a perfect recall?", it seems much more apparent now that our preferences would surely change, had we had more information, less computational cost for processing that information, and less bias from our memories. Looking back at all these interesting findings, it would seem apparent that our brains are highly complex, interconnected and highly dependent on the simultaneous functioning of their subsystems. This sounds, almost as if, there is no established subsystem in the brain which

embodies the full will of the system, nor are there notions which hint at that possibility (Chalmers, 2018).

4.4 Preferences: Brain Modularity

If there were a subsystem within the brain that is independent of other subsystems that dominates the ordering of preferences, thus the decision-making processes, then predicting behaviour and future preferences might be feasible. However, recent neural evidence suggests otherwise. If our brains were closed systems, then reliably predicting behaviour with accuracy and consistency is possible. However, our brains are open systems, where each subsystem is reliant on and interacts with multiple subsystems simultaneously. If this is the true case of how our brains function, then it is not possible to pin down any specific subsystem or function as one which dominates the ordering of preferences and, in turn, our decision-making processes. In other words, when we develop preferences and act on those preferences by making choices, multiple subsystems within the brain influence these choices. Although new brain imaging technologies such as fMRI may be able to provide us with knowledge regarding our tendencies, however, this would be without revealing a specific underlying mechanism that produces an act of choice (Martins, 2007a; 2011b).

For example, the hippocampus is activated when engaging in explicit value recall processes; the amygdala is related to encoding some classes of negative outcomes; the orbitofrontal cortex encodes immediate sensory-derived value for options under consideration; the anterior cingulate cortex is related to the activity of tracking the cost of opportunities (I.e. tradeoffs); the dorsal striatum is correlated to the action of storing values that are continuously encountered (I.e. habituation or stabilizing preferences); the anterior insular cortex regards the formulation of

values correlated with aversiveness; the medial prefrontal cortex and ventral striatum correlate with subjective value; finally, the dorsolateral prefrontal cortex appears to be highly correlated in exhibiting patience and self-control (I.e. discounting the future). All of those regions are correlated to value and preference-related behaviours, and while these regions may aid us in predicting choice in specific areas, there is no single region responsible for the formulation of preferences that lead to those specific choice actions; no specific region stands out (Bartra et. al., 2013, as cited in Wilson and Kirman 2016; 89).

Similarly, Symmonds and Dolan (2011) show that there appears to be no specific region within the brain of which we can observe the specific formulation of our preferences and how they change over time. Rather, the function of our preferences is not region-dominated but it is, in fact, the notion of interconnectedness itself between all of the brain regions or subsystems. While there are specific regions in the brain that mediate specific functions attributable to risk attitudes, or rather, regions that become activated (Gianotti, Knoch, Faber et al., 2009, as cited in Sharot, 2011), however, all brain regions need to coordinate simultaneously to produce preference-directed behaviour, no specific region is directly correlated with preferences, as earlier research attempted to reveal. In other words, our preferences can be observed as connectivity between multiple regions or subsystems; it is connectivity itself that is the driving force of preferences, rather than having a single or multiple subsystems that are functionally independent of other regions. This can perhaps mean that the only way to truly understand how preferences function in the brain is by simulating the whole brain and observing the strength of connectivity between all of the brain regions. If this is the case, then perhaps we are quite far from truly capturing what our preferences really are.

In this chapter, I investigate three key neural aspects related to how our preferences formulate and adapt throughout time: memories, choice sets, and brain modularity. Firstly, it appears that our preferences are greatly dependent on our memories. However, our memories are greatly amenable to change, deviations from the truth, and completely false memory constructions. If our preferences are predicated to a large extent on what we remember from past events, where our memories contain partial or complete false reconstruction, then our preferences may not be reliably predicted, at least in individual cases. This is the case, because much of our memories escape conscious thought and deliberation, thus making it much more challenging for us to understand how our present preferences were formulated and how they could further deviate or cohere in the future. Secondly, various studies on the size of choice sets and choice blindness reveal interesting information about our preferences. For example, the ways in which our preferences are formulated cannot escape the notion of computational cost at times when we make decisions and observe various choice sets. While computational cost may be irrelevant to an artificial agent, it is quite relevant to how our preferences adapt over time. Furthermore, choice blindness alludes to the idea that our preferences are highly malleable, and that we can be quite resistant to our own recollections and interpretations of our preferences in the future.

Lastly, studies on brain modularity reveal that there is no independent seat in the brain which takes full or the majority of control of our preferences. Rather, our preferences arise from the connectivity between different brain regions, and not from any specific or multiple regions that are independent from other brain regions. This means that, if there are no specific markers within the brain which are directly related to the encoding and later revealing of our preferences, then it may become more challenging to develop a general theoretical framework that fully captures our preferences. Additionally, this entails that we need to revisit the concept of preferences, and

reassess whether and how preferences should be defined differently than or more similarly to values.

Chapter Five: Analysis

This paper investigates the transdisciplinary conceptual case study of preferences as one which lies at the center of the AI value alignment problem as an emerging security issue. The analysis incorporates theory from various approaches to human preferences, such as biological evolution, mainstream economics, and neurocognitive studies. Each discipline contributed tremendous knowledge which aids the current understanding of our preferences. However, there remains no general theoretical framework on preferences that can reconcile the differences and similarities in findings from those disciplines. Considering that the AI value alignment problem represents a hypothetical future instance, this paper lacks means of empirical experimentation and validation. Thus, this research is reliant on theory and concept in design, and extrapolates this from the available literature from primary and secondary sources, while specifically focusing on the key theories deemed necessary in flushing out the problem.

5.1 *Discussion*

This paper reaches two key conclusions about our preferences which can help inform the current work on AI value alignment problems as an emerging security issue. The findings engage Russell's (2019; 176) propositions for developing a provably beneficial AI:

1) *The machine's only objective is to maximize the realization of human preferences, with human behavior being the ultimate source of information about those preferences*

Standard economics theory proposes that we are inherent maximizers, meaning that we constantly try to opt for the optimal choice, and this tendency is reflected in our revealed preferences (Samuelson, 1948). When incorporating other frameworks into the analysis of

preferences, such as biological evolution and neuroscience, it becomes apparent that maximization is indeed an intrinsic feature of the human condition which can be observed from our preferences and overall behaviour (Burnham and Phelan, 2019; Dolan and Sharot, 2011). For example, the biological evolutionary framework is specifically beneficial to understanding whether our preferences for maximization are predicated upon some formal mechanistic processes (e.g. disposition for specific genetic markers). At the beginning of this paper, I suggested that if there are formal mechanisms showcasing specific patterns underlining our preferences, and those patterns are tractable, then perhaps they can be replicated and modelled algorithmically. If tractability is achieved, then we can build a general theoretical framework upon which AI can more reliably capture our true preferences and infer our values from them. If AI cannot reliably capture our preferences and what we value, then we run the risk of having developed a highly intelligent AI that can threaten our security if it cannot understand what we want and do not want.

One interesting idea emerges when looking at the literature of maximization. Just as in standard economics theory, where it is assumed that the individual maximizes some utility "as if" she was rational, the same reasoning is applied in AI research on inferring human values from observing human preferences "as if" we were maximizing some known utility. It is assumed that, if the AI can capture our true preferences, then it would be able to capture what we truly value (Russell, 2019). Our preferences reflect the image of us behaving "as if" we were efficient maximizers. Here, I pose an alternative question. Rather than basing our analysis of preferences on the notion of maximizing "as if", we should instead redirect this notion to an "even if" problem which deserves equal attention. Even if we were able to efficiently maximize any utility (e.g. happiness, wealth, fame) would that necessarily represent reliable and stable predictions of our values?

Would maximization here, even if it were to reflect some absolute truth about us, be beneficial when realized on a mass scale? To further elaborate, assume this scenario. Annie's preferences in academia are reflecting her value for academic success and recognition. Annie is a perfectly rational and efficient maximizer, as she succeeds in maximizing all utilities deemed relevant to her success. However, upon graduating the top of her class in a high regarded university, Annie realizes that she is unhappy and rather unsatisfied. Does this mean that she did not, in fact, value her academic success? Or that there exists something which she values more than this?

AI research adopts Reinforcement Learning (RL) and Inverse Reinforcement Learning (IRL) methods, which were derived from behavioural psychology, in devising ever more intelligent algorithms that can gradually become more efficient at inferring our values from our preferences (Russell, 2019). However, one aspect of these two techniques appears to be neglected in AI research: Post Reinforcement Pause (PRP). This refers to the idea that an individual, over time, will come to value the preceding effort that is put into maximization over the final value or reward which is derived at the completion of the objective (Griffiths and Thompson, 2017). In other words, perhaps Annie had exercised too much effort into valuing her academic success, and once she had attained this success, its value had already lost some or perhaps all of its worth.

Our preferences do reflect what we may or may not value, that is true. However, our preferences also reflect a strong correlation between maximization and satisfaction, where the more effort put into maximization the less satisfaction we derive from the reward. The more we tend to maximize, the less happy we are (Schwartz, 2004). Neurocognitive research affirms this notion, by studying the various regions within the brain which become activated when engaging in problem solving in decision making. It appears that we tend to prefer to navigate through less information to reach a 'good enough' decision, rather than navigate a ton of information to reach

the optimal decision. In a sense, maximization is embedded within the behaviour of satisficing, rather than in the behaviour of maximization itself.

It is reasonable to neglect coding the concept of PRP into a machine, given that the machine does not exhibit any notion of subjectivity or satisfaction. Additionally, the machine is much more computationally capable when compared to us, thus it would not be exercising effort in the same sense that we do. However, the problem here is that, even if AI can reliably infer our values from our preferences, and can take away much of the effort we usually tend to put into choosing between competing options, this does not necessarily mean that an AI that arrives at the optimal choice on our behalf would ensure our satisfaction. Firstly, just as we tend to ignore too much information when making a choice, we can reasonably assume that we might also tend to ignore too much information provided by the AI when asked how it reached a certain optimal choice on our behalf. In other words, we may blindly trust the AIs strategy by virtue of it being smarter and faster than us in conducting competing calculations, and in the process of doing this, the AI might still be making false inferences about our values, thus consolidating and indirectly reaffirming our hidden biases and preferences which might not, in fact, maximize our wellbeing in the long run.

Additionally, while coding a variant of PRP into a machine appears unnecessary for the machine in terms of reaching optimality, PRP is, however, crucial to how our preferences adapt throughout time and thus affect our current and future values. Annie's experience of PRP may prompt a change in her future preferences in terms of wanting to achieve more balance between optimization and subjective satisfaction. Thus, where Annie had initially thought she valued academic success, with experiencing PRP, she realizes that what she truly values is something different than this. Thus, her future preferences and values would adhere to this realization accordingly.

Furthermore, I argue that as long as AI is incapable of accessing our memories, which appear to largely constitute our present and future preferences, it will remain incapable of reliable comprehension and prediction of our preferences. Thus, the challenge of inferring our values remains persistent. The literature on our memories is tremendous, all showcasing their malleability (Tekin and Roediger, 2017). This malleability appears to be crucial in our preference formulation, where even the slightest adjustment of past events and future recall can incur resistance and drastic preference change for the individual. For example, merely being asked about our intentions regarding having made a specific choice in accordance with a preference ordering for that choice, the question causes us to change or reverse our preferences. In false recollection experiments on memory, subjects tended to display resistance to later recounts of the same memory they initially had at the beginning of the experiment (Johannsson et. al., 2008).

Thus, it appears that more comprehensive research on the cognitive processes constituting our memories is needed in order to build AI that is better capable at making inferences about our preferences and values. In a sense, this means that not only would AI need to infer our values from our preferences, but also infer the degree to which our memories are shaping our various preference orderings. Effectively, this means that the AI would need to infer our memories alongside its inferences about our preferences. Human behaviour reveals a good deal of information about human preferences, however, much of what constitutes our preferences lies beyond what we can consciously and immediately access. Human memories, subliminal cues in the environment, genetic markers and social interaction, all contribute to the underpinnings of our preferences. Human behaviour is merely the source which can be directly observed, however, it is not the *ultimate* source through which our preferences can be understood and reliably captured.

The second axiom of standard economics proposes, while considering some informational constraint and being able to make voluntary choices, that we nevertheless succeed in maximization. However, we tend to rather fail at this. Standard economics theory conditions this failure upon either being forced to make a choice, or lacking information that is detrimental for making the optimal choice. When looking at failures of maximization through multiple lenses from outside the field of standard economics, one concludes that the above two presuppositions are rather insufficient in explaining maximization failure. People tend to fail at maximization even when they make voluntary choices and have sufficient information. For example, a drug addict may be very well aware of the negative consequences and cost of drug addiction that will be incurred in the long run, yet nevertheless remains actively seeking procurement of drugs.

One particular mechanism which can explain maximization failure and irrational behaviours and preferences, is one which biological evolution can offer us. Evolutionary scientists incorporate evolutionary mismatch as a potential explanation to this. Evolutionary mismatch can explain how our preferences have become maladaptive, by tracing this maladaptiveness from the agricultural revolution, well into the first industrial explosion and beyond. The idea is that genes which determine certain behavioural inclinations have failed to catch up to the quick and drastic change in our environment, rendering some of our genetic predispositions as outdated. According to this line of reasoning, our preferences had been consistent and rational in our ancestral times, yet diverged into inconsistency and irrationality due to the technological growth which had happened too fast for natural selection to work its way into adjusting our preferences accordingly. If evolutionary mismatch is empirically representative of a causative chain, through which the changes in our preferences can be reliably traced, then this may have serious implications in our current understanding of our preferences. For example, ASI can pose yet

another evolutionary mismatch in our timeline, perhaps one that is greater than the preceding industrial events in our history by virtue of ASI being the last and greatest invention of humanity. If our preferences diverge and become inconsistent when we are not in equilibrium with the environment, then ASI might tip the current balance even further than before, producing a more extreme version of divergence and inconsistency.

2) The machine is initially uncertain about what those preferences are

As opposed to the earlier approaches on the standard AI model, Russell (2019; 187-213) argues that uncertainty about the objective is key to developing a provably beneficial AI because uncertainty would prevent the AI from maximizing arbitrary and dangerous objectives that do not correspond to our intentions. Additionally, an uncertain AI would allow us to turn it off, which is a key element in ensuring AI safety. The idea is that, as long as the AI is uncertain about our true preferences, then it would always allocate some probability to all attributes which may actually contribute to our preferences. Thus, allowing itself to be turned off may very well be within the range of preference probabilities. In reference to the notion that as AI gets sufficiently more intelligent and capable, it would become more certain about our preferences, it might become dangerous if it reaches certainty by having made false initial inferences about our preferences. However, Russell (2019) further argues that even as the AI becomes more intelligent, as long as it always allocates some probability to other potential preferences then it will remain "perfectly manageable" (p. 204).

Russell (2019) discusses certainty only through two potential arguments. Firstly, certainty can be benign, where the AI becomes more certain about our true preferences, i.e. it becomes more 'right' about us. The second kind of certainty is less benign, where the AI becomes more certain about false preferences which it had been incorrectly inferring at the beginning of the process.

This type of certainty resembles the outdated standard model of AI that maximizes objectives which we do not desire.

I argue that there could be a third kind of certainty, one that is both true about our preferences yet at the same time non-benign. While the concept of uncertainty appears to be crucial in developing a safe AI, it appears that the idea of there being formal markers which determine our preferences can be, in contrast to what I suggested earlier in the paper, potentially dangerous and pose a new version of a security dilemma. It is believed that if we can build a general theoretical framework on preferences, thereby truly capturing our preferences and reliably trace their formulation and underpinnings, then we can build AI that would learn how to better infer our values from our preferences. However, this notion seems to contrast the crucial element of uncertainty which appears to be detrimental for ensuring AI safety. AI can only remain uncertain about us insofar as there exists no formal and fundamental truth about our preferences. However, if there is such a truth, then the AI is no longer uncertain, and it may begin a process of dangerous maximization towards that formal truth. This problem sounds almost paradoxical, as it entails that we may need to reconcile uncertainty with formal truth, which are two key contrasting concepts.

For example, in reference to our biological evolution, and specifically the concept of evolutionary mismatch, if there is a causative framework that explains the current divergence of our preferences, and as AI gets more intelligent it figures it out, then it may start maximizing our preferences according to this causal chain which represents a formal truth about the human condition. This means maximizing “as if” we were optimal, according to a far point in our evolutionary history which was viable then but is not viable in the same sense anymore. The problem, as I mentioned earlier, is an “even if” problem. Even if our preferences are predicated

upon hardwired biological or genetic patterns (e.g. survival and reproduction) in a provably systematic way, we may not want this AI to make choices that accord or compliment those outdated hardwired patterns, even if they contain some kind of absolute truth about us. The idea is that, even if AI gets our preferences 'right', it may only choose to optimize our lives according to preferences orderings that would only ensure the long-term biological survival and increased reproduction, effectively ignoring our subjective experiences of those preferences. Humans may, on a collective scale, share a common preference ordering, while not being aware of its importance, or aware of it at all. However, conditional on becoming aware of this hardwired preference ordering could be that we become highly resistant of collective convergence according to that preference ordering as AI attempts to steer us towards it.

For example, an AI of this sort may conclude that the best way to ensure it maximizes our truly provable preferences (e.g. survival and increased reproduction) would be by eradicating any and all objects and situations which may endanger the realization of these preferences. Such things could range from taking full control of technologically operated weapons or even destroying them, and eradicating jobs which contain some amount of risk, which could be any kind of job. Effectively, this kind of AI may end up entirely changing our landscape to prevent us from realizing any preference ordering that may counteract the preference ordering which contains the formal truth.

There seems to be a subtle yet potentially an important connection between the theoretical frameworks regarding human preferences as suggested by proponents of various disciplines. Schwartz (2004) refers to a persistent conflict between maximizing and satisficing. Kahneman (2011) refers to a persistent conflict between our remembering self and our experiencing self (I.e. system one and system two). Mlodinow (2012) refers to the conflict between our conscious

and our unconscious mind. Burnham and Phelan (2002) refer to this conflict as arising from evolutionary mismatch contrasted with willpower. When one looks at the explanations of preference inconsistencies arising from those various disciplines, one cannot help but notice a pervasive dialectical pattern reflecting our dual nature: there is a driving force in us which seeks to put us on a formal maximization path, yet there is another force that constantly collides with the former one, producing irrationalities and inconsistencies. I argue that, while both of the aforementioned driving forces hold truth about our preferences, a provably beneficial AI might decide to realize the preferences of only one but not the other. This scenario represents a hypothetical info-hazard dilemma which can arise from the maximum realization of one force at the expense of the other, where one holds a formal truth and the other engulfs all of the uncertainty and 'randomness' constituting our behaviours.

I will refer to the maximization aspect of the human condition as the *formal* one by virtue of it being fundamental, provable, and systematic. On the other hand, the inconsistent aspect will be referred to as *non-formal*. The former contains certainty or "proof" about our preferences which could, upon its realization, set the AI on a dangerous path of optimizing a life which we do not desire, thereby threatening our security and safety. On the other hand, the latter appears to be the one through which we can ensure that the AI can be turned off, and would not optimize an entirely different landscape of preferences of which we would not approve of by virtue of this path containing uncertainty. A provably beneficial AI might need to complement our non-formal preferences, as opposed to the formal one. In other words, in order for us to prevent the AI from reaching certainty (even if certainty here reflects non-debatable truth) we must first ensure that there is no formal truth about our preferences which this AI can capture once it becomes sufficiently intelligent. If there is such a formal truth about us, one which binds all of our values

under a single unit of preference ordering that is rational for this AI, and it maximizes it, our current landscape might become greatly unstable and unpredictable. Furthermore, our resistance to this maximization process might -in turn- incur various security threats and endanger our safety.

5.2 *Conclusion*

This paper explores our preferences as its main conceptual unit of analysis, with the aim of unravelling the existence of formal markers that underpin our preferences and characterizes them in a reliable and predictive way. The study of our preferences is crucial in AI value alignment research as an emerging security problem about a hypothetical future instance in which AI would become vastly superior to us in all domains, thus we must ensure that it is aligned with our values. Said values can be inferred from our preferences, thus making preferences an important element to solving the value alignment problem. The theoretical analysis on preferences incorporates theory from mainstream and behavioural economics, biological evolution, and neurocognitive research. The reliance on the latter two disciplines is justified as potentially being able to provide us with empirical results and tractability on our preferences. If we can fully capture our true preferences, then we can develop AI that can better infer our values from our preferences, thereby mitigating the security threats which AI can incur if it were not aligned with our values.

This paper reaches two key conclusions. One negates the idea that our behaviour is the ultimate source of our preferences, by referring to numerous studies which reveal subliminal and biological factors which constitute our preferences, which are not observable through behaviour. AI would need to infer more than our preferences from our observable behaviour in order for it to correctly capture our values. The second conclusion reveals a potential security dilemma between formal markers constituting our preferences and the notion of uncertainty as being key

to developing a provably beneficial AI. The idea of formal truth about our preferences collides with uncertainty, meaning that if our preferences hold a formal and provable mathematical truth, then the AI need not remain uncertain about us as it becomes sufficiently more intelligent and capable. The more certain AI is about our preferences, the higher the probability of producing security and safety concerns by becoming misaligned to what we value. Additionally, with the rising certainty, it becomes much more challenging for us to control or contain the AI, thereby referring to the crucial problem of not being able to turn it off.

It is generally assumed that as long as AI can truly understand us, our intentions and our values, then things will be manageable and the AI will not pose major security threats. However, this paper shows that this is not necessarily true. It appears that even if AI truly captures the human condition, then we are still in some kind of danger as this represents a security dilemma that deals with the concept of info-hazard. An AI that truly understands us is not necessarily a beneficial AI that will ensure our safety in the manner in which we intend, if knowledge about the human condition contains a formal absolute truth which can only become dangerous upon being realized by this AI. Our conflicting dialectical nature alludes to this idea, and the more that AI learns about us could mean the more dangerous this AI becomes, if it cannot achieve the same kind of balance which we have achieved, if it decides to maximize only the aspect about us which contains the formal truth.

It might be beneficial for future research on AI value alignment to focus on ensuring robustness between uncertainty and what could be considered as observable or truly non-observable truths about our preferences. Furthermore, it would be helpful to reassess whether AI inferring our values from our preferences is enough, or if it is necessary to incorporate more than our revealed preferences and direct instructions in building this kind of AI. This paper showed that our

revealed preferences are only representative of partial aspects of our values, and that other crucial aspects about our values may not be immediately accessible through our observable preferences and even conscious deliberate thought. It is believed that truly intelligent machines can achieve our objectives as we intend, only if they fully capture what we want and what we value. However, now it appears that the concept of uncertainty is potentially a crucial one to maintain in developing an intelligent AI. Yet, to fully capture our desires and values while remaining indefinitely uncertain about us appears to be a contradictory process.

A potential research question can investigate the implications of formal truths and uncertainty, and whether the two concepts can be reconciled and maintained in AI that will become vastly superior to us. In conclusion, even if there is a fundamental formal truth about our preferences, that may not necessarily be a solution to the value alignment problem. Rather, this may represent an even more difficult problem which must be further explored and studied as an emerging security dilemma in the literature on AI safety and security. If uncertainty is key to the solution here, then any framework which seeks to bind our preferences and values under a single collective preference ordering would cancel the safeguarding effect of uncertainty, which would mean that the security threats of a misaligned AI would nevertheless persist.

References

Abe, N. et. al. (2008). Neural Correlates of True Memory, False Memory, and Deception. In *Cerebral Cortex*, 18(12): (pp. 2811-2819). https://doi.org/10.1093/cercor/bhn037

Albrecht, S. V. and Stone, P. (2018). Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems. In *the Journal of Artificial Intelligence*, 258: (pp. 66-95). https://doi.org/10.1016/j.artint.2018.01.002

Arnemann, K. L., Chen, A. J., Novakovic-Agopian, T., Gratton, C., Nomura, E. M., & D'Esposito, M. (2015). Functional Brain Network Modularity Predicts Response to Cognitive Training after Brain Injury. *Neurology*, *84*(15), 1568–1574. https://doi.org/10.1212/WNL.0000000000001476

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge, UK: Cambridge University Press.

Bostrom, N. (2018a). 'Strategic Implications of Openness in AI Development', in R, Yampolskiy (ed.) *Artificial Intelligence Safety and Security*: Boca Raton, FL: CRC Press/Taylor & Francis Group.

____ 2014b. *Superintelligence: Paths, Dangers and Strategies*, Oxford University Press.

Bortolotti, L. (2004). Can We Interpret Irrational Behavior? In *Behavior and Philosophy,* 32(2).

Brown, G. R. and Richerson, P. (2014). Applying Evolutionary Theory to Human Behaviour: Past Differences and Current Debates. In the *Journal of Bioeconomics* 16: (pp. 105-128).

Burnham, T.C., Phelan, J. (2019a). Ordinaries. In the *Journal of Bioeconomics,* 21: (pp. 145–155). Available at https://doi.org/10.1007/s10818-019-09287-z

____ 2000b. *Mean Genes: From Sex to Money to Food, Taming our Primal Instincts*. Cambridge, Mass: Perseus Pub.

Burnham, T. C. (2016a). Economics and Evolutionary Mismatch: Humans in Novel Settings do not Maximize. In the *Journal of Bioeconomics* 18: (pp. 195–209). Available at https://doi.org/10.1007/s10818-016-9233-8

____ 2013b, Towards A Neo-Darwinian Synthesis Of Neoclassical And Behavioral Economics. *Journal of Economic Behavior & Organization*. 90: (pp. 113-127). Available at https://doi.org/10.1016/j.jebo.2012.12.015

Burnham, T.C., Dunlap, A.S., and Stephens, D. (2015). Experimental Evolution and Economics. *SAGE Open, 5*, 2158244015612524.

Chalmers, D. J. (2018a). The Meta-Problem of Consciousness. In *the Journal of Consciousness Studies* 25: (pp. 6-61).

____ 2012b. *Constructing the World*. Oxford University Press.

____ 2010c. The Singularity: A Philosophical Analysis. In the *Journal of Consciousness* (17) 9-10.

____ 1995d. Facing Up to the Problem of Consciousness. In *the Journal of Consciousness Studies*, 2(3).

Capra C.M., and Rubin P.H. (2011). Rationality and Utility: Economics and Evolutionary Psychology. In: Saad G. (eds) *Evolutionary Psychology in the Business Sciences*. Springer, Berlin, Heidelberg.

Ceci, S. J. et al. (1994). Repeatedly Thinking About Non-Events. In *the Journal of Consciousness and Cognition*, 3: (pp. 388-407). DOI:10.1006/ccog.1994.1022

Cesarini, D., Johannesson, M., Lichtenstein, P., Sandewall, Ö, & Wallace, B. (2010). Genetic Variation in Financial Decision-Making. In *The Journal of Finance, 65*(5): (pp. 1725-1754). Available at www.jstor.org/stable/40864985.

Cesarini, D. (2009). Genetic Variation in Preferences for Giving and Risk Taking. *Quarterly Journal of Economics*, 124(2): (pp. 809-842). Available at http://hdl.handle.net/1721.1/57904

Chaiken, S., and Trope, Y. (1999). *Dual-Process Theories in Social Psychology*. New York: Guilford Press.

Coolidge, F. (2020). *Evolutionary Neuropsychology: An Introduction to the Evolution of the Structures and Functions of the Human Brain*. Oxford University Press.

Cubitt, R. P., and Sugden, R. (1998). The Selection of Preferences Through Imitation. *Review of Economic Studies*, Oxford University Press, 65(4): (pp. 761-771).

Omohundro, S. M. (2008). The Basic AI Drives. In *Proceedings of the 2008 Conference on Artificial General Intelligence: Proceedings of the First AGI Conference*. IOS Press, NLD: (pp. 483-492).

Damasio et. al. (2008). Differential Effects of Insular and Ventromedial Prefrontal Cortex Lesions on Risky Decision-Making. *Brain: A Journal of Neurology*, *131*(5): (pp. 1311-1322). https://doi.org/10.1093/brain/awn066

Damasio, A. R. (2004). '*Emotions and Feelings: A Neurobiological Perspective.*' In A. S. R. Manstead, N. Frijda, & A. Fischer (Eds.), *Studies in Emotion and Social Interaction. Feelings and Emotions: The Amsterdam symposium*. Cambridge University Press. (pp. 49–57). https://doi.org/10.1017/CBO9780511806582.004

DeepMind [website], accessed 1 May 2020, https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii

Demski, A. and Garrabrant, S. (2019). *Embedded Agency*. arXiv:1902.09469v1 [cs.AI]

Demski, Abram. (2018). *Decision Theory*. Machine Intelligence Research Institute (MIRI), available at https://intelligence.org/2018/10/31/embedded-decisions/

Dennett, D. C. (1991). *Consciousness Explained*. United Kingdom: Penguin Books.

Dolan, R. J., and Sharot T. (2011). *Neuroscience of Preference and Choice: Cognitive and Neural Mechanisms.* London: Elsevier/Academic Press.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics, 75*(4). DOI: 10.2307/1884324

Evans, O., Stuhlmueller, A., and Goodman, N. D. (2015). *Learning the Preferences of Ignorant, Inconsistent Agents.* arXiv:1512.05832v1

Facebook AI [website], accessed 27 April 2020, https://ai.facebook.com/blog/facebook-leads-wmt-translation-competition/

Finley, J.S. et. al. (2017). Relative Contributions of Semantic and Phonological Associates to Over-Additive False Recall in Hybrid DRM Lists. In the *Journal of Memory and Language*, 93: (pp. 154-168).

Freedman, D. J. et al. (2001). Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. In *the Journal of Science*, 291(5502): (pp. 312-316). DOI: 10.1126/science.291.5502.312

Gass S.I., and Harris C.M. (2001). *Combinatorial Explosion*. In: Gass S.I., Harris C.M. (eds) Encyclopedia of Operations Research and Management Science. Springer, New York, NY.

Good, I. J. (1965). *Speculations Concerning the First Ultraintelligent Machine*, In Advances in Computers, Franz L. Alt and Morris Rubinoff (Eds.), 6: New York: Academic Press.

Gowdy, J., Rosser, B. and Roy, L. (2013), The Evolution of Hyperbolic Discounting: Implications for Truly Social Valuation of the Future. In the *Journal of Economic Behavior & Organization*, 90, (S): (pp. 94-104). Available at https://doi.org/10.1016/j.jebo.2012.12.013

Greenwood, J. and Seshadri, A. (2002). *The U.S. Demographic Transition*. FRB of Cleveland Working. (pp. 01-18). Available at http://dx.doi.org/10.2139/ssrn.297952

Griffiths, R.R., and Thompson, T. (1973). The Post-Reinforcement Pause: A Misnomer. *The Psychological Record*, 23: (pp. 229–235). https://doi.org/10.1007/BF03394160

Harbi et. al., (2015). Efficiency, Equality, Positionality: What do People Maximize? Experimental vs. Hypothetical Evidence from Tunisia. In the *Journal of Economic Psychology*, 47. Available at https://doi.org/10.1016/j.joep.2015.01.007

Heyes, C. (2010). Where Do Mirror Neurons Come From? In *the Journal for Neuroscience and Biobehavioral Reviews, 34*(4): (pp. 575–583). Available at https://doi.org/10.1016/j.neubiorev.2009.11.007

Hintze, A., Olson, R.S., Adami, C., & Hertwig, R. (2015). Risk Sensitivity as an Evolutionary Adaptation. *Scientific Reports, 5*. DOI:10.1038/srep08242

Hoeffler, S. (2002). Constructing Stable Preferences: A Look Into Dimensions of Experience and Their Impact on Preference Stability. In the *Journal of Consumer Psychology*, (8). Available at https://doi.org/10.1207/s15327663jcp0802_01

Irons, W. (1998). Adaptively Relevant Environments Versus the Environment of Evolutionary Adaptedness. In the *Journal of Evolutionary Anthropology*, 6: (pp. 194-204). Available at https://doi.org/10.1002/(SICI)1520-6505(1998)6:6<194::AID-EVAN2>3.0.CO;2-B

Johansson, P. H., and Sikström, S. L. (2008). From Change Blindness to Choice Blindness. *Psychologia*. 51 (2). Available at https://doi.org/10.2117/psysoc.2008.142

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., and Tversky, A. (1992a). Advances in Prospect Theory: Cumulative Representation of Uncertainty. In *the Journal of Risk Uncertainty,* 5. Available at https://doi.org/10.1007/BF00122574

____ 1981b. The Framing of Decisions and the Psychology of Choice. In the *Journal of Science, 211*(4481): (pp. 453-458). Available at www.jstor.org/stable/1685855

____ 1979c. Prospect Theory: An Analysis of Decision under Risk. In the *Journal of Econometric Society*. 47 (2): (pp. 263–291). DOI: 10.2307/1914185

Kelkar A. and Medaglia J.D. (2018). Evidence of Brain Modularity. In: Shackelford T., Weekes-Shackelford V. (eds) *Encyclopedia of Evolutionary Psychological Science*. Springer, Cham. https://doi.org/10.1007/978-3-319-16999-6_2422-1

Keysers, C. & Gazzola, V. (2006). *Towards a Unifying Neural Theory of Social Cognition*. Progress in Brain Research, 156: (pp. 379-401). DOI: 10.1016/S0079-6123(06)56021-2

Kriegeskorte, N., Douglas, P.K. (2018). *Cognitive* Computational Neuroscience. *Nat Neuroscience,* 21: (pp. 1148-1160). Available at https://doi.org/10.1038/s41593-018-0210-5

Lakshminarayanan, V. and Santos, L. R. (2011). The Evolution of Our Preferences: Insights from Non-human Primates. In R. J. Dolan & T .Sharot (eds). *Neuroscience of Preference and Choice: Cognitive and Neural Mechanisms.* London: Elsevier/Academic Press. (pp. 75-89).

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1997). 'Motivated attention: Affect, activation, and action'. In P. J. Lang, R. F. Simons, & M. T. Balaban (Eds.), *Attention and Orienting: Sensory and Motivational Processes*. Lawrence Erlbaum Associates Publishers. (pp. 97–135).

Levy, M. (2010). Evolution of Risk Aversion: The 'Having Descendants Forever' Approach. SSRN Electronic Journal. Available at http://dx.doi.org/10.2139/ssrn.1688265

Leisman G. et. al. (2012). Intentionality and "Free-Will" from a Neurodevelopmental Perspective. In *the Journal of Frontiers in Integrative Neuroscience*, 6. Available at https://doi.org/10.3389/fnint.2012.00036

Lima, S. (1984). Downy Woodpecker Foraging Behavior: Efficient Sampling in Simple Stochastic Environments. In the *Journal of Ecology, 65* (1): (pp. 166-174). Available at https://doi:10.2307/1939468

Loftus, E. (2005). Planting Misinformation in the Human Mind: A 30-Year Investigation of the Malleability of Memory. In *Learning & Memory,* 12(4): (pp. 361-6) (Cold Spring Harbor, N.Y.). DOI: 10.1101/lm.94705

Martins, N. (2011a). Can Neuroscience Inform Economics? Rationality, Emotions, and Preference Formation. *Cambridge Journal of Economics*. 35(2): (pp. 251-267).

____ 2007b. Realism, Universalism, and Capabilities. *Review of Social Economy*, 65(3): (pp. 253-278).

Mlodinow, L. (2012). *Subliminal: How Your Unconscious Mind Rules Your Behavior*. New York: Pantheon Books.

Morwitz, V. G., and Fitzsimons, G. J. (2004). *T*he Mere-Measurement Effect: Why Does Measuring Intentions Change Actual Behavior? *Journal of Consumer Psychology, 14*(1-2). Available at https://doi.org/10.1207/s15327663jcp1401&2_8

Naselaris, T. et. al. (2018). Cognitive Computational Neuroscience: A New Conference for an Emerging Discipline. *Trends in Cognitive Sciences*, *22*(5). Available at https://doi.org/10.1016/j.tics.2018.02.008

Neisser, U., and Fivush, R. (Eds.). (1994). *T*he Remembering Self: Construction and Accuracy in the Self-Narrative. Emory Symposia in Cognition, 6. Cambridge University Press. https://doi.org/10.1017/CBO9780511752858

Nickerson, R. S. and Adams, M. J. (1979). Long-Term Memory for a Common Object. In the *Journal for Cognitive Psychology*, 11.

OpenAI [website], accessed 12 June, 2020, https://openai.com/blog/openai-api/

OpenAI [website], accessed 29 April 2020, https://openai.com/projects/five/

Owain E., Andreas S., and Noah D. G., (2015), *Learning the Preferences of Ignorant, Inconsistent Agents*. arXiv:1512.05832v1 [cs.AI]

Peterson, M. (2006). Indeterminate Preferences. Philosophical Studies: *An International Journal for Philosophy in the Analytic Tradition, 130*(2): (pp. 297-320). Available at www.jstor.org/stable/4321801

Premack, D. (2010). Why Humans Are Unique: Three Theories. *Perspectives on Psychological Science*, 5(1): 22–32. Available at https://doi.org/10.1177/1745691609356782

Rayo, L. and Becker, G. (2007). Evolutionary Efficiency and Happiness. In the *Journal of Political Economy*, 115(2): (pp. 302-337). Available at https://doi.org/10.1086/516737

Roediger, H. L. (1974). Inhibiting Effects of Recall. In *the Journal of Memory & Cognition,* 2: (pp. 261-269).

Roediger, H. L., and Neely, J. H. (1982). Retrieval Blocks in Episodic and Semantic Memory. *Canadian Journal of Psychology,* 36: (pp. 213-242).

Roediger, H. L., and Gallo, D. A. (2004). 'Associative Memory Illusions'. In R. F. Pohl (Ed.), *Cognitive illusions: A Handbook on Fallacies and Biases in Thinking, Judgment and Memory.* Oxford: Oxford University Press. (pp. 309-326).

Rubin, P. and Paul, C. (1979). An Evolutionary Model of Taste for Risk. In the *Journal of Economic Inquiry*, 17(4): (pp. 585-96). Available at https://doi.org/10.1111/j.1465-7295.1979.tb00549.x

Rueschemeyer, D. (2009). *Usable Theory: Analytic Tools for Social and Political Research*. Princeton; Oxford: Princeton University Press. (pp. 87-106). doi:10.2307/j.ctvcm4j6k

Russel, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Random House LLC.

Sotala. K. (2018). 'Disjunctive Scenarios of Catastrophic AI Risk' in R, Yampolskiy (ed.) *Artificial Intelligence Safety and Security*: Boca Raton, FL: CRC Press/Taylor & Francis Group.

Sotala, K. and Gloor, L. (2017). *Superintelligence as a Cause or Cure for Risks of Astronomical Suffering*. In *Informatica*, Foundational Research Institute.

Samuelson, P. (1948). Consumption Theory in Terms of Revealed Preference. *Economica, 15*(60): (pp. 243-253). DOI: 10.2307/2549561

Schacter D. L. (2012). Constructive Memory: Past and Future. Dialogues in *Clinical Neuroscience*, *14*(1): (pp. 7–18).

Schwartz, B. (2004). *The Paradox Of Choice: Why More Is Less*. New York: Ecco.

Searle, J. R. (1989c). *Artificial Intelligence and the Chinese Room: An Exchange*. *New York Review of Books*, 36: 2.

____ 1984b. *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

____ 1980c. *Minds, Brains, and Programs*. *Behavioral and Brain Sciences*, (3): (pp. 417-57).

Simon, H. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69: (pp. 99-118).

Skinner, B. F. (1953). *Science and Human Behavior.* New York: Macmillan.

Soares, N. & Fallenstein, B. (2017). *Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda*. 10.1007/978-3-662-54033-6_5.

Sporer, S.L. (2008). Lessons from the Origins of Eyewitness Testimony Research in Europe. In the *Journal of Applied Cognitive Psychology*, 22. doi:10.1002/acp.1479

Sporns, O., & Betzel, R. F. (2016). Modular Brain Networks. *Annual Review of Psychology*, *67*, 613–640. https://doi.org/10.1146/annurev-psych-122414-033634

Thaler, R. H. (1987). Anomalies: The January Effect. In the *Journal of Economic Perspectives*, 1 (1): (pp. 197-201). Available at https://doi:10.1257/jep.1.1.197

Thaler, R. H., and Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Alfred A. Knopf.

Tekin, E., and Roediger, H. (2017). The Range of Confidence Scales Does Not Affect the Relationship Between Confidence and Accuracy in Recognition Memory. *In Cognitive Research: Principles and Implications*. 2. 10.1186/s41235-017-0086-z.

Toshiba Corporation. (2019). Toshiba's Breakthrough Algorithm Realizes World's Fastest, Largest-scale Combinatorial Optimization. In *PhysOrg*, accessed 27 April 2020, https://phys.org/news/2019-04-toshiba-breakthrough-algorithm-world-fastest.html

Wade, K. A. et al. (2002). A Picture Is Worth a Thousand Lies: Using False Photographs to Create False Childhood Memories. In *Psychonomic Bulletin & Review* 9(3).

Wallace, C. S. and Dowe, D. L. (1991). Minimum Message Length and Kolmogorov Complexity. In *the Computer Journal*, vol. 42(4): (p. 270). Oxford University Press.

Weintraub .E. R. (2007). *Neoclassical Economics*. The Concise Encyclopedia Of Economics.

Wells, G. & Olson, E. (2003). Adult Eyewitness Testimony. *Annual Review of Psychology*. 54: (pp. 277-95). 10.1146/annurev.psych.54.101601.145028.

Wilson, D. S., & Kirman, A. (2016). *Complexity and Evolution: Toward a New Synthesis for Economics*. Cambridge, Massachusetts; London, England: MIT Press.

Witt, U. (2016). The Transformations of Utility Theory: A Behavioral Perspective. In the *Journal of Bioeconomics* 18: (pp. 211–228). Available at https://doi.org/10.1007/s10818-016-9235-6

Woodworth, B. et. al. (2018). *Preference Learning in Assistive Robotics: Observational Repeated Inverse Reinforcement Learning*. Proceedings of Machine Learning Research, 85: (pp. 1-19).

Wu Y. et. al. (2016). *Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation, available at arXiv:1609.08144 [cs.CL]*

Yudkowsky, E. and Soares, N. (2018). *Functional Decision Theory: A New Theory of Instrumental Rationality*, available at arXiv:1710.05060 [cs.AI]

Yudkowsky, E. (2016). *The AI Alignment Problem: Why it's Hard and Where to Start*, Machine Intelligence Research Institute (MIRI). Available at https://intelligence.org/files/AlignmentHardStart.pdf

Yudkowsky, E. and Herreshoff, M. (2013). *Tiling Agents for Self-Modifying AI, and the Löbian Obstacle*. Machine Intelligence Research Institute (MIRI). Available at https://intelligence.org/files/TilingAgentsDraft.pdf