Evolutionary perspectives on AI values

Maria Avramidou

Faculty of Philosophy, University of Oxford,
Radcliffe Humanities, Woodstock Road, Oxford OX2 6GG, UK

This was a prize-winning entry into the Essay Competition on the Automation of Wisdom and Philosophy.

INTRODUCTION

Artificial Intelligence (AI) possesses the potential to develop capabilities that, if misaligned with human values, could present an existential risk to humanity (Bostrom, 2014; Russell, 2019). This threat is fundamentally rooted in the alignment problem—the challenge of ensuring that AI systems' goals align with human values. This problem can be subdivided into three primary issues: i) the value specification problem, which pertains to precisely defining human values in a format that an AI can interpret and act upon without deviation; ii) the value monitoring problem, which involves ensuring that an AI system genuinely possesses the correct values rather than merely appearing to do so; and iii) the value drift problem, which concerns developing AI in such a manner that its objectives remain aligned with human values over time.

As we consider the future of automation, the decision to automate AI will hinge on our assessment of its reliability. If we reach a point where we can sufficiently trust AI, we might be inclined to automate a wide range of tasks, even high-stakes ones like political decision-making. Conversely, if trust remains an issue, we might limit AI automation to less critical tasks and maintain human oversight for significant decisions. Thus far, discussions have predominantly focused on whether AI can learn the right values, and how we can determine when it fails to do so. It is crucial to develop a robust framework for understanding how values evolve over time.

The introduction of the theory of evolution by natural selection (Darwin, 1859) revolutionised our understanding of the development of life on Earth. According to the theory, species experience random mutations that create variations in traits affecting their survival. Some organisms are better suited to their environment according to selection pressures–external factors that determine which traits are advantageous and thereby influence the fitness of an entity. These organisms have an evolutionary advantage and are, thereby, more likely to pass on their traits to the next generation. These traits become increasingly prevalent within a population. Over time, this process results in the evolution of species that are increasingly well-adapted to their environments.

For an organism to be subject to Darwinian evolution, the following conditions must be met: i) there must be differences in traits within a population. These variations may arise from mutations or other sources, providing the raw material for selection; ii) these traits must be inheritable, meaning they can be passed down from one generation to the next. This ensures that

advantageous traits can be transmitted to offspring; iii) external factors influence which traits are advantageous, gradually shaping the species.

Dawkins (1983) coined the term *Universal Darwinism,* recognising the application of the core Darwinian principles beyond biological phenomena. The idea is that any system with entities that vary, replicate, and face selection pressures undergo a form of evolution. This concept suggests that Darwinian principles can explain the development of diverse systems, from cultural practices to economic behaviours. Early precedents of the idea of generalising Darwinism to evolving systems outside biology include, among others, applications to political institutions (Bagehot, 1872; Ritchie, 1891), economics (Hodgson, 1996), human learning and the development of science (James, 1880), ethical principles (Kidd, 1894), and social evolution (Veblen, 1899; Baldwin, 1909; Keller, 1915).

Just as the principles underlying Darwinian evolution can explain the development of life, they too can help us understand AI development. In particular, these principles apply to the kinds of values AI systems may develop and how these values might evolve over time. AI models undergo a form of "evolution", where AI engineers—or even the algorithms themselves—develop various models to identify the most effective one for performing a given task or set of tasks. In this sense, AI models go through iterative cycles of "mutations" and "fitness tests". Those models that perform best are kept and further refined, while less effective ones are discarded.

Depending on the tasks they are assigned to perform, models might develop different values. When AI models are trained for different tasks, they end up valuing different things based on what they're designed to optimise for. Thus, when predicting how the values of AI systems might evolve over time, we should consider the incentives present in the environment that might lead AIs to adopt certain values. For example, if AI systems are used more for some set of tasks, they may build up more training data for these tasks, shaping their view of the world across different tasks. Preferences associated with these tasks might be disproportionately amplified. Therefore, the values represented globally will change significantly based on how these systems are predominantly used. This matters because if AI systems are replicated and widely deployed for various tasks, their values and preferences will shape—and potentially dominate—the environments and societies they interact with. Thus, the evolution of AI values will be driven by the ways in which these systems are deployed and the environments they adapt to. The literature

on value drift aims to argue for the inevitability of selfishness in a way that will be detrimental to humans, by drawing from economic models and examples of competition in nature.[1]

In this essay, I explain how Universal Darwinism applies to AI systems and argue that it does not suggest that AIs will necessarily evolve to become selfish, despite claims made in the literature. In Section I, I begin with an overview of the different aspects of the alignment problem. In Section II, I present how evolutionary theory applies to AI systems. In Section III, I explore the development and evolution of AI values.

SECTION I: THE ALIGNMENT PROBLEM

Value specification

Translating abstract human values into precise, actionable guidelines that can be understood and implemented by an AI system involves defining these values as well as prioritising them in situations where they might conflict. Due to the fact that there is not a universally accepted value system, determining which values should guide AI behaviour is one of the bottlenecks to AI alignment. In particular, human values are diverse and can differ significantly not just across cultures but even among individuals within a given culture. What is considered moral, ethical, or important can vary widely, making it challenging to program these nuances into an AI system. The difficulty lies not only in the initial encoding of these values but also in ensuring that the AI's interpretation aligns with human intentions across diverse and unforeseen scenarios. Values are often context-dependent. Consequently, actions considered acceptable in one situation may be deemed inappropriate in another. This context-dependency adds an extra level of complexity to the task of defining values in a way that AI can consistently apply them.

Approaches to overcome this bottleneck include imitation learning, where the AI learns to imitate behaviour through observation; inverse reinforcement learning (IRL), where the AI learns to model human values through behavioural observations; reinforcement learning from human feedback (RLHF), where the AI refines its actions based on direct feedback from humans (Christiano et al., 2017); iterated amplification and distillation (IDA), where the AI incrementally improves its decision-making by repeatedly breaking down complex tasks into simpler sub-tasks and refining its performance through multiple iterations (Christiano et al., 2018).[2]

[1] For instance, see Hendrycks (2023).

[2] For a more extensive discussion on these approaches and their limitations, see Dale and Saad (2024).

Value monitoring

A significant challenge that arises from the methodologies employed in AI development is ensuring that AI systems truly embody the desired values, rather than merely dissembling. The need for verification mechanisms is critical, as these mechanisms must be able to detect discrepancies between the AI's declared values and its actual operational objectives, thus preventing deceptive alignment.

One approach to tackle this issue is mechanistic interpretability, which involves reverse-engineering the computational mechanisms and representations learned by neural networks into human-understandable algorithms and concepts to provide a granular, causal understanding of human values and intentions (Olah, 2023).

Value drift

Just as human values can evolve, so too might the values of AIs, particularly as they experience new environments or integrate new information. The integration of digital minds into broader digital or mixed environments that involve interactions with other digital minds as well as humans will result in changes in what they prioritise and value. AI systems could be able to learn and evolve in ways that may stray from their initial alignment and the divergence of an AI's operational objectives from human values could lead to unintended and potentially disastrous consequences.

This issue becomes increasingly perilous given the capabilities that AI systems will likely possess to operate at much higher speeds. While biological Darwinism appeals to random mutations over a long period, AI development is guided by human intervention on a radically compressed timescale. What might be centuries of "thinking" and "learning" time for an AI could equate to merely minutes in our subjective human experience. During these periods, there could be significant shifts in values. Humanity itself has seen profound shifts in values over the centuries; consider the transformation in societal norms and moral perspectives from the Middle Ages to the modern era, which saw shifts from feudalistic values to modern democratic and human rights values. Although such changes have generally contributed to the well-being of humanity, the outcome of rapid value changes in AI does not necessarily promise the same positive trajectory.

For instance, rapid shifts in ethical norms in AI could lead to undesirable outcomes if the new values adopted by the AI are misaligned with human welfare.

The faster drift rate in AI systems necessitates the creation of robust mechanisms not just for initial value alignment but also for ongoing monitoring and re-calibration of values. As AI evolves, these mechanisms must be capable of adjusting or reinforcing the AI's values in real-time to prevent misalignment that could occur due to rapid value evolution. Guidelines would be needed to manage how values are updated or reconciled with societal norms. However, monitoring is constrained by interpretability methods, as effective oversight depends on the ability to not only observe but also verify that the AI's actions and decision-making processes remain consistent with its intended values.

Additionally, alignment measures themselves could lead to value drift. For example, methods like reinforcement learning from human feedback, intended to refine an AI's actions and values through human interaction, might inadvertently shift an AI's value system through the introduction of biases present in the feedback. The compounding effect of continuous feedback over accelerated periods of AI operational time could lead to a significant and unintended drift in values.

There has surprisingly been little attention given to the problem of value drift in the literature, relative to the other problems.[3] However, understanding the mechanism of value drift appears crucial for predicting and influencing the short- as well as long-term behaviour of AI.

SECTION II: DARWINIAN EVOLUTION OF AI SYSTEMS

In this section, I explain how Darwinian evolution applies to both biological organisms and AI models. I clarify the statistical interpretation of evolution and the role of competition and death. I also identify the units of selection and beneficiaries in the context of the evolution of AI systems.

Statistical evolution

Despite the misconceptions that resulted from the title of Dawkins' "The Selfish Gene" (Dawkins, 1976), evolution is not guided by one specific goal, such as the survival of the fittest

[3] For discussions on the other problems, see for instance Rabinowitz et al. (2018); Amodei et al. (2016); Doshi-Velez and Kim (2017); Orseau and Armstrong (2016); Bostrom (2014).

genes or creation of more complex organisms. Such teleological interpretations miss the point. Darwinian evolution is not tied to the specifics of genes, DNA, or complexity; instead, it is a framework that explains how the environment shapes existence and frequency of traits.

Darwinian evolution is the process by which entities best suited to their environment are more likely to survive and reproduce. A "fitter" entity, at any given time, has a higher probability of survival. And the longer it lives, the more opportunities it has to pass on its traits. In physical organisms, this translates to producing offspring that share, amongst all traits, the advantageous ones. For example, if having blue eyes somehow made a person better adapted to their environment, that person would have more chances to reproduce. Assuming that the choice to reproduce is uniformly distributed within the population—meaning an equal percentage of people with and without blue eyes choose to reproduce—the frequency of blue-eyed individuals would increase due to their longer survival. Changes in the frequency of traits within a population are driven by this differential reproductive success (Walsh et al., 2002), with traits that confer higher fitness becoming more prevalent over time. Meanwhile, individuals lacking advantageous traits are less likely to survive and reproduce, gradually reducing the frequency of those traits in the population. Traits that confer a reproductive advantage statistically tend to increase in frequency over generations, not because nature intrinsically "selects for them", but because organisms with these traits are more likely to pass them on successfully.

In the context of AI, models that perform exceptionally well at a given task will tend to be preferred over others because they offer more accurate, efficient, or reliable results. This preference leads to increased use and further refinement of these successful models. As they are adopted more widely, they receive more resources, including computational power. Conversely, less effective models are used less frequently and may eventually be completely discarded. This cycle results in a Darwinian evolution of AI models, where only the most effective ones thrive.

Extinction

None of this implies that a reduction in the absolute number of individuals in a population or the extinction of species with an unfit trait is a ubiquitous aspect of the evolutionary framework. Darwinian evolution refers to the change in the frequency of traits. Of course, we notice that some species survive while others die off, largely because species have finite timelines. If species

lived indefinitely–and assuming they did not have to compete for the same resources–the numbers could simply increase for both the fittest and the unfit members.

Entities die off either because they lack the necessary resources to survive, they are eliminated by other entities, or due to randomness. In humans, ageing and poor fitness are examples of lack of necessary resources, whereas competition with other entities for resources or malicious intent—even from less fit entities—are instances of elimination by other entities. On the other hand, car accidents are an example of a random event. For AI models, "death" occurs when they are no longer run. This can happen because they become obsolete due to the emergence of superior, more effective models. If computational resources are limited, humans will allocate fewer and fewer resources to less effective models and eventually replace them completely. A model might also be discarded because it no longer meets user needs. Even if no other models exist, if a model fails to perform a task at the desired level of performance, it will not be used. Accidental death of AI models on the other hand could be an accidental deletion without retrieval.

Units of selection

When considering how evolutionary pressures act on different entities, there are two main questions to address: which entities are subject to Darwinian evolution and who benefits long-term from selection. I provide an answer to these questions in turn in the context of AI models and human users.

Subject to Darwinian evolution are all entities that "replicate", in the sense of by passing on their structure largely unchanged, without necessarily exhibiting material overlap across generations (Hull, 1980). Instead, generations are linked by informational relationships (Godfrey-Smith, 2009). This notion extends the definition of a replicator beyond biological organisms. Replication involves only the copying of a property, be it a gene, an idea (a meme), a piece of information, or a structure (Nanay, 2011). Just as biological entities replicate genetic material and cultures propagate memes, AI models can be copied for use by different users and purposes. This capacity for replication and modification means that AI systems, too, undergo a form of Darwinian evolution where successful adaptations are retained and unfavourable ones are discarded.

Humans and AIs as beneficiaries

Beyond the entities that are subject to Darwinian evolution, there are entities that benefit from it. Entities can be either primary or secondary beneficiaries. That is, they can directly benefit from their own fitness, gaining an advantage in survival and reproductive success from their adaptive traits. At the same time, they may benefit indirectly from the fitness of others. For instance, ecosystems might gain stability from the successful adaptations of key species within them. The evolutionary success of certain plant species can bolster entire ecosystems by providing essential resources to a wide range of organisms. In a world where humans and AIs coexist, humans can be secondary beneficiaries, benefiting from the fitness of AIs. For instance, the capabilities of AI systems can lead to technological advancements that benefit human society in myriad ways. However, this benefit is contingent upon the alignment of AI evolution with human values.

Additionally, it is often the case that mutual benefits between different individuals, groups, or species arise when their interactions lead to advantages on both sides. Consider for instance the relationship between flowering plants and their pollinators. Bees collect nectar and pollen from flowers as food. In doing so, they inadvertently carry pollen from one flower to another, facilitating cross-pollination, and hence, enhance the fitness of flowers by making them more resilient to environmental changes Such mutual benefit could also be present between humans and AIs. For instance, AI systems often improve their algorithms and increase their capabilities by learning from human input. Machine learning models used in applications like language translation, image recognition, or recommendation systems refine their accuracy by analysing large datasets generated from human interactions. By processing and learning from human-generated data, AI systems can optimise their algorithms, become more efficient at their tasks, and thereby be preferred by humans over other, less optimal, AI systems. At the same time, humans could benefit from AI systems that provide more accurate, personalised, and efficient services. However, as AIs gain more capabilities, they need not depend on human data for learning. AI systems might have access to other data streams beyond those directly generated by human interactions, including those generated by other AI systems. Additionally, with the development of robotics and multimodal systems, there will be a continuous stream of data from interactions with the world. I will next explore whether the strategy that maximises fitness for AIs and humans will be a cooperative one. This will be the focus for the remainder of this essay.

SECTION III: AI VALUES THROUGH THE EVOLUTIONARY LENS

Values are not static; they evolve over time as the environment changes. Values such as cooperation, altruism, and fairness have likely been shaped because they increase the chances of survival within a social context. How the environment rewards or penalises different traits determines which different sets of traits correspond to higher fitness in that environment. For AIs, the selection pressures will arise from their operational and functional environments—essentially, the tasks they are designed to perform and the contexts in which they operate. An AI's programmed values might drift when subjected to external pressures such as changes in operational demands, human interaction, or shifts in societal norms. Even an aligned AI may become misaligned as societal norms in digital ecosystems evolve or as different incentives are exercised. Addressing these questions provides a framework for predicting and managing value drift in AI systems and guiding design choices.

Applications, performance, and values

Considering selfishness as "the lack of consideration for others and concern for one's own personal profit", we can explore whether evolutionary pressures will necessarily promote selfish traits. Are more effective models always more selfish? This section explores the relationship between effectiveness and selfishness, examining whether optimal performance necessitates self-centred behaviour or if cooperative traits can also be favoured by evolution.

Using specific examples from nature to justify an argument can result in circular reasoning. The mere fact that certain traits exist in nature does not necessarily make those examples relevant. The fittest traits—physical or personal—are uniquely defined for each specific context at any given moment. To answer the question, we need to consider what traits will be rewarded, rather than assuming selfishness in ways that will harm humans simply because we can create simplified toy models based on animals in nature. There has been a marked trend toward altruism. An illustrative example of this is observed in bee colonies, where worker bees, which are sterile, dedicate their lives to protecting the hive and supporting the reproductive queen, a behaviour that benefits the colony's survival as a whole (Wilson, 1971). These cooperative strategies improve the fitness of the group without implying selfishness.

The question of whether AIs will be willing to sacrifice themselves for the benefit of a group of AIs, or a combined group of AIs and humans, depends on the incentives and selection pressures

they face. A model does not necessarily need to be selfish toward humans to perform well on a task. While some tasks may benefit from selfishness—for example, robbing a bank involves prioritising personal gain over others—peak performance does not imply that a model should disregard human well-being. In the context of AI, a model that excels at a task may do so because it has been designed to learn effectively from data, adapt to user needs, or collaborate with other systems. This improved performance is not a sign of selfishness but rather an indication of effectiveness in achieving desired outcomes.

The applications of AI shape the values they adopt and exhibit. If AI systems are predominantly used in tasks that encourage selfish behaviours—such as adversarial strategies—they may reinforce those values culturally. The values promoted by the majority of AI systems will become more prevalent and influential within human society. Those values will have a larger representation in society and will therefore shape behaviours and norms by increasing the visibility of those values through interactions. Conversely, deploying AI in areas where optimal performance requires cooperation--like collaborative problem-solving--fosters and amplifies cooperative values. By doing so, we can guide the influence of AI toward a more collaborative direction.

Benevolence towards less capable entities

Hendrycks (2023) argues that *"Group selection only makes sense when a group is more effective than some subset of the group breaking off. Unless humans add value to a group of AIs, AI-human groups would fail to outcompete groups composed purely of AIs. AIs would likely do better by forming their own groups."* In the context of evolutionary theory, "more effective" can be thought of as "possessing higher fitness".

Thus, Hendrycks suggests that additional members that do not add value reduce the fitness of the group. This seems odd. For if the presence of less useful members does not adversely affect performance on the task at hand, the fitness need not be reduced. Hence, those members could remain in the group.

For analogy, consider the role of flowers in human lives. Humans plant and care for flowers not because they directly contribute to our survival but because of aesthetic preferences. Specifically, while humans who avoided things that hindered their survival had a higher probability of living longer and reproducing, our responses to entities that neither positively nor negatively impact our

survival can vary. Sometimes we may eliminate these entities, but in other cases, we develop norms that promote their preservation.

Thus, we should ensure that we instil foundational norms in AIs to protect humans while we are still useful to them. The question remains whether they will replace these norms once they realise they are arbitrary. This topic is treated–albeit with a slightly different focus–in discussions on how AIs will be susceptible to philosophical arguments, such as evolutionary debunking arguments.[4]

Cooperation in times of scarcity

Hendrycks also suggests that *"AIs are more likely to see one another as part of their group, so they will tend to be cooperative with one another and competitive with us."* However, since group formation is not necessarily based on relatedness, but rather on fitness, AIs will not necessarily be more cooperative towards each other compared to humans.

In scenarios where resources are scarce, there is a trade-off between effectiveness and resource intensity. AIs may have greater capabilities than humans, but they also require more resources. It might, therefore, be more advantageous for AIs to eliminate humans and other AIs to free up resources. While it might seem rational for AIs to compete against each other for resources, cooperative strategies could actually yield higher overall efficiency. Cooperative AIs could develop systems of resource sharing and task specialisation that maximise the use of available resources far more effectively than if each AI acted alone. By reducing conflict and pooling capabilities, cooperative groups of AIs might achieve greater accomplishments and maintain a more stable existence than solitary or adversarial AIs.

Most importantly, humans can play a critical role in the early competition between groups of AIs, potentially becoming integral members of a human-AI coalition. This scenario hinges on humans having something valuable to offer in this competition with AIs, which is expected to be real-world resources in the stage before AIs are widely embodied. Before AIs achieve full autonomy and embodiment, their dependence on human-provided infrastructure, energy, and computational resources could form the basis for mutually beneficial relationships. This mutual dependency could also ensure that AI developments are steered towards outcomes that are advantageous not only for AIs but also for humanity.

[4] See for instance Dale and Saad (2024).

Although we currently lack a robust framework for predicting value change, we possess effective heuristics for incentivising different behaviours that may not have originally aligned with their values but now do, due to strong incentives. For example, economic incentives have been shown to motivate environmental conservation behaviours among individuals who previously prioritised economic gain over ecological preservation (Pirard, 2012).

If AIs inevitably evolve selfish values, it suggests a significant risk in allowing them to take on high-stakes tasks without stringent oversight and control. Therefore, to mitigate these risks, it would be prudent for humans to ensure that AI automation remains constrained to low-stakes tasks. This approach would involve restricting AI deployment to areas where the potential impact of selfish behaviour is minimal and manageable, thereby reducing the likelihood of adverse consequences.

CONCLUSION

The evolution of AI values is a crucial factor in determining the extent of its automation. Evolutionary theory suggests that, just as biological organisms have adapted over millennia through natural selection, AIs too will undergo evolutionary changes shaped by digital environments and interaction pressures.

The potential for rapid value drift in AIs—a consequence of their ability to process and evolve at rates far surpassing human cognition—poses significant safety risks. Such shifts could lead to misalignment with human values, potentially resulting in outcomes that are detrimental to societal well-being.

The interaction between humans and AIs introduces additional layers of complexity when thinking about how AI values will evolve. Humans stand to benefit enormously from AI through enhancements in efficiency, problem-solving capabilities, and the management of complex systems. Conversely, AIs can learn from human creativity and enhance their own functionalities. This reciprocal relationship underscores the potential for a symbiotic evolution, where AIs and humans co-evolve, potentially leading to new forms of cooperative strategies that are beneficial to both. The role of humans in this evolving landscape is both as contributors and beneficiaries. By instilling foundational norms and values in AI systems and fostering an environment that

encourages beneficial AI behaviours, humans can guide AI evolution in directions that augment rather than undermine human interests. This proactive approach will require not only technological and scientific expertise but also a deep philosophical and ethical engagement with the implications of AI in society.

ACKNOWLEDGEMENTS

Thanks to Bradson Saad and Gustavs Zilgalvis for their helpful comments.

REFERENCES

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). *Concrete problems in AI safety*. arXiv preprint arXiv:1606.06565.

Bagehot, W. (1872). *Physics and Politics, or, Thoughts on the Application of the Principles of "Natural Selection" and "Inheritance" to Political Society*. London: Henry King.

Baldwin, J. M. (1909). *Darwin and the Humanities*. Baltimore: Review Publishing.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press, UK.

Christiano, P., Shlegeris, B., and Amodei, D. (2018). *Supervising strong learners by amplifying weak experts*. arXiv preprint arXiv:1810.08575.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). *Deep reinforcement learning from human preferences*. Advances in neural information processing systems, 30.

Dale, M. T. and Saad, B. (2024). *Evolutionary debunking and value alignment.* Global Priorities Institute Working Paper Series, No. 11-2024.

Darwin, C. (1859). *On the origin of species by means of natural selection.* London: Murray Google Scholar.

Dawkins, R. (1983). *Universal Darwinism*. In Evolution from Molecules to Man, edited by D. S. Bendall. Cambridge University Press, UK.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press, UK.

Doshi-Velez, F. and Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv preprint arXiv:1702.08608.

Godfrey-Smith, P. (2009). *Darwinian populations and natural selection.* Oxford University Press, USA.

Hanson, R. (2016). *The age of Em: Work, love, and life when robots rule the earth*. Oxford University Press.

Hendrycks, D. (2023). *Natural selection favors AIs over humans*. arXiv preprint arXiv:2303.16200

Hodgson, G. M. (1996). *Economics and evolution: bringing life back into economics.* University of Michigan Press.

Hull, D. L. (1980). *Individuality and selection.* Annual review of ecology and systematics, 11(1):311–332.

James, W. (1897). *Great Men, Great Thoughts, and the Environment.* In William James, *The Will to Believe and Other Essays in Popular Philosophy*, 216–54. New York and London: Longmans Green. Originally published in the Atlantic Monthly (vol. 46, 1880, 441–59).

Keller, A. G. (1915). *Societal Evolution: A Study of the Evolutionary Basis of the Science of Society*. New York: Macmillan.

Kidd, Benjamin. (1894). *Social Evolution*. London and New York: Macmillan.

Nanay, B. (2011). *Replication without replicators*. Synthese, 179:455–477.

Olah, C. (2023). *Interpretability dreams.* Informal notes.

Orseau, L. and Armstrong, M. (2016). *Safely interruptible agents*. In Conference on Uncertainty in Artificial Intelligence . Association for Uncertainty in Artificial Intelligence.

Pirard, R. (2012). *Market-based instruments for biodiversity and ecosystem services: A lexicon.* Environmental science & policy, 19:59–68.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., and Botvinick, M. (2018). *Machine theory of mind*. In International conference on machine learning , pages 4218–4227. PMLR.

Ritchie, D. G. (1891). *Darwinism and politics: With two additional essays on human Evolution* (Vol. 4). S. Sonnenschein & Company, lim.

Russell, S. (2019). Human compatible: AI and the problem of control. Penguin, UK.

Smith, J. M. and Price, G. R. (1973). The logic of animal conflict. Nature, 246(5427):15–18.

Veblen, T. B. (1899). *The Theory of the Leisure Class: An Economic Study in the Evolution of Institutions*. New York: Macmillan.

Walsh, D. M., Lewens, T., and Ariew, A. (2002). *The trials of life: Natural selection and random drift.* Philosophy of Science, 69(3):452–473.

Wilson, E. O. (1971). *The insect societies*. Harvard University Press, USA.