Will AI Avoid Exploitation?

Adam Bales (Global Priorities Institute, University of Oxford)

Global Priorities Institute | December 2023

GPI Working Paper No . 16-2023

Please cite this working paper as: Bales, A. (2023) Will AI Avoid Exploitation? *Global Priorities Institute Working Paper Series, No. 16-2023* Available at: https://globalprioritiesinstitute.org/adam-bales-will -ai-avoid-exploitation

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Will AI Avoid Exploitation?[1]

Artificial General Intelligence and Expected Utility Theory

Adam Bales

1 Introduction

Recent decades have seen rapid progress in artificial intelligence (AI). Some people expect that in the coming decades, further progress will lead to the development of AI systems that are at least as cognitively capable as humans (see Zhang et al., 2022). Call such systems artificial general intelligences (AGIs). If we develop AGI then humanity will come to share the Earth with agents that are as cognitively sophisticated as we are.[2]

Even in the abstract, this seems like a momentous event: while the analogy is imperfect, the development of AGI would have some similarity to the encountering of an intelligent alien species who intend to make the Earth their home. Less abstractly, it has been argued that AGI could have profound economic implications, impacting growth, employment and inequality (Korinek & Juelfs, Forthcoming; Trammell & Korinek, 2020). And it has been argued that AGI could bring with it risks, including those arising from

[1] This paper is forthcoming in *Philosophical Studies*. For feedback, thanks to Jacob Barrett, Hilary Greaves, Rose Hadshar, Andreas Mogensen, Christian Tarsney, Teru Thomas, Elliott Thornley, Hayden Wilkinson, and Timothy L. Williamson.

[2] For my purposes, little rests on whether there's a sense in which AGIs won't count as agents. As long as AGIs can carry out cognitive tasks to a sufficiently high level then it's worth exploring the implications of such systems.

human misuse of powerful AI systems (Brundage et al., 2018; Dafoe, 2018) and those arising more directly from the AI systems themselves (Bostrom, 2014; Carlsmith, Forthcoming).

Given the potential stakes, it would be desirable to have some sense of what AGIs will be like if we develop them. Knowing this might help us prepare for a world where such systems are present. Unfortunately, it's difficult to speculate with confidence about what hypothetical future AI systems will be like.

However, a surprisingly simple argument suggests we can make predictions about the behaviour of AGIs (this argument is inspired by Omohundro, 2007, 2008; Yudkowsky, 2019).[3] According to this argument, we should expect AGIs to behave as if maximising expected utility (EU).

In rough terms, the argument claims that unless an agent decides by maximising EU it will be possible to offer them a series of trades that leads to a guaranteed loss of some valued thing (an agent that's susceptible to such trades is said to be *exploitable*). Sufficiently sophisticated systems are unlikely to be exploitable, as exploitability plausibly interferes with acting competently, and sophisticated systems are likely to act competently. So, the argument concludes, sophisticated systems are likely to be EU maximisers. I'll call this the *EU argument*.

In this paper, I'll discuss this argument in detail. In doing so, I'll have four aims. First, I'll show that the EU argument fails. Second, I'll show that reflecting on this failure is instructive: such reflection points us towards more nuanced and plausible alternative arguments. Third, the nature of these more nuanced arguments will highlight the limitations of our models of AGI, in a way that encourages us to adopt a pluralistic approach. And fourth, reflecting on such models will suggest that at least sometimes what matters is less developing a formal model of an AGI's decision-making procedure and more clarifying what sort of goals, if any, an AGI is likely to develop. So while my discussion will focus on the EU argument, I'll conclude with more general lessons about modelling AGI.

2 AGI, What and When?

I start with background on artificial general intelligences.

Above, I said that AGIs are AI systems that are at least as cognitively capable as humans. Despite what AGI stands for, this definition doesn't rely on any unified notion of

[3] Yudkowsky's views are more nuanced than is suggested by the EU argument. I'll point to some relevant nuances later.

general intelligence, but is instead compatible with the idea that there are simply a set of distinct cognitive skills. This might lead to vagueness in some comparisons of cognitive capability, but in other cases the differences across these skills will be decisive enough for one being to be reasonably described as more cognitively capable than another (as is the case with humans and chickens). We could then think of AGIs as systems that are more cognitively capable than humans in this sort of decisive way. Indeed, we could weaken this definition: we'll be in uncharted territory if AI systems become able to carry out creative, scientific, economic, political and military cognitive tasks in a way that's even roughly comparable to humans. I'll assume this weaker notion in what follows.

I said that some people expect AGI to be developed in the coming decades. More concretely, in one survey of machine learning experts, the aggregated forecast assigned a 50% chance to human-level AI systems being developed by 2060 (Zhang et al., 2022). Another survey assigned a 50% chance to (greater than) human-level AI by 2059 (Stein-Perlman et al., 2022). And a model that draws on comparisons to the human brain projects a 50% chance of "transformative" AI by 2052, where AI is transformative if its impact is on the same scale as the industrial revolution (Cotra, 2020).[4] Overall, these considerations suggest a serious possibility that AGI is developed within half a century.

Still, I doubt we should lean heavily on these estimates. As to the surveys, I'm sceptical that experts in machine learning are also experts in predicting the future progress of machine learning; predicting future trends in AI is simply a different skill to implementing current ML techniques. As to Cotra's model, I worry that such models reveal less about the future than about the modelling assumptions made, for example about the connection between the computing power likely to be applied in the future to training AI and the likely performance of future systems. Overall, I'm far from confident that we'll develop AGI by 2060.

Still, I think that humanity's history with technology suggests a tendency to first make radical advances and only afterwards consider the profound impacts these technologies could have. For example, while the first nuclear bomb was dropped in 1945, a sound understanding of the possibility of nuclear winter developed only slowly over decades, during which time the number of nuclear warheads rose steadily to a peak of about 64,000 in 1986 (Roser et al., 2013). That's a lot of warheads to produce without understanding the potential implications of their use. Likewise with fossil fuels and a sophisticated

[4] Transformative AI needn't necessarily be AGI. Still, expectations about when we'll develop transformative AI are informative about when we should expect to develop AGI.

understanding of climate change. Contra these cases, it seems desirable that we understand the risks a technology poses *before* deploying it en masse. So, given rapid recent advancements in AI, and given expert predictions, and given that progress sometimes takes us by surprise, I think AGI is worth reflecting on now, even if it's unclear when, if ever, such systems will be developed.

3 Expected Utility Theory

Turning to background on expected utility (EU) theory, consider a bet on a fair coin where you win £10 on heads but lose £4 on tails. This bet has an expected monetary value of £3, which is found by weighing each payoff by the chance of receiving it (0.5*10 + 0.5 *-4 = 3). More generally, we can assign expected monetary values to any gamble if we know the monetary payoffs and the chances of these payoffs.

EU theory is an attempt to use similar reasoning to provide a general theory of decision making. According to this theory, the expected value of a decision is a probability-weighted sum of the value of the decision's possible outcomes. However, these values can't be monetary, as monetary payoffs aren't the only thing that matters in evaluating decisions. Instead, the values are *utilities*, where these provide a (cardinal) representation of how desirable an agent finds an outcome. And if we're interested in decisions, we don't want to focus on objective chances, as these are typically unknown and so unhelpful for decision making. Instead, we focus on *credences*, where these are probabilities representing the agent's degree of belief that each outcome will result.[5]

To spell this out formally, let S represent the set of possible states of the world, each of which captures one (possibly coarse-grained) way the world could be. Let Cr represent the agent's credence function and U represent the agent's utility function. Then the EU of an act, *a*, is defined as:[6]

$$EU(a) = \sum_{s \in S} Cr(s)U(sa)$$

That is, to calculate the EU of a decision you assign utilities to possible outcomes of your decision, where outcome *sa* captures what happens in the state, *s*, given that the act, *a*,

[5] We could instead use the credences that the agent *ought* to have given their evidence. This distinction is unimportant for my discussion.

[6] I set aside the debate around evidential and causal decision theory (Peterson, 2017, ch. 9), which is largely independent of the current issue, although Garrabrant (2022) suggests that the EU argument collapses if we adopt a particular competitor to both theories.

is carried out. You then sum these utilities, after weighting them by the credence in the relevant state.

The claim that AGI will decide by maximising EU could be interpreted as the claim that these systems will have an explicit credence and utility function, will calculate the EUs of the various acts they could carry out, and will choose the act that maximises EU. Alternatively, the claim could be that AGI will act *as if* this were so. This second version of the claim takes no position on the internal structure of these systems; instead, it simply states that the systems can be modelled as if deciding by maximising EU. For purposes of predicting systems, it doesn't matter which of these claims is true, so I'll focus on the latter (weaker) claim.

4 The EU Argument

Underpinning EU theory are mathematical results called representation theorems. According to these theorems, if an agent's preferences satisfy certain axioms then the agent will act *as if* they're maximising EU. That is, their behaviour will be able to be modelled as if they have some credence and utility function and always make the decision that maximises EU according to these functions.

To spell this out more slowly, let a *prospect* be a probability distribution over possible outcomes (where these probabilities capture the agent's credences).[7] For example, the prospect associated with leaving your umbrella at home might assign probability 0.2 to you remaining dry (because the rain might hold off) and probability 0.8 to you getting wet (because it might rain). Then let $\succ$ represent strict preference over prospects and $\succcurlyeq$ represent weak preference (that is, $A \succcurlyeq B$ indicates that either $A \succ B$ or the agent is indifferent between the prospects). These preferences should be interpreted behaviourally, as specifying how the agent would be disposed to choose if given a choice between A and B.[8] A representation theorem can then proceed by appeal to the following axioms:[9]

[7] Throughout this paper, I'll take for granted that it's reasonable to talk of an agent's credences, though this is itself a substantial assumption that could be challenged.
[8] I'm not committing myself to behaviourism, as my interest isn't in ascribing mental states to AI systems but instead in predicting these systems' behaviour (for relevant discussion, see Clarke, 2016). Similarly, I'm not taking a general stance on the revealed preferences approach to EU theory but am simply considering an approach that's potentially fruitful in the current context.
[9] Fishburn, 1970. I restrict my focus to prospects with a finite number of possible outcomes.

1. *Completeness.* For all A and B, either $A \succcurlyeq B$ or $B \succcurlyeq A$ (or both). Informally, completeness states that you can compare all prospects because either you (strictly) prefer one to the other or you're indifferent between them.
2. *Transitivity.* For all A, B, and C, if $A \succcurlyeq B \succcurlyeq C$ then $A \succcurlyeq C$. We're familiar with transitivity in other contexts. For example, if Alice is taller than Bob and Bob is taller than Carol then Alice is taller than Carol. The transitivity axiom requires that (weak) preference between prospects have this same structural feature.
3. *Continuity.* For all A, B, and C, if $A \succ B \succ C$ then there are non-trivial p and q, such that $pA+(1-p)C \succ B \succ qA+(1-q)C$, where $pX+qY$ indicates a prospect that has a probability p of resulting in X and a probability q of resulting in Y. So continuity says that if you have a middling preference for B then you will be willing to forego B and risk C for a high enough chance of A, but will prefer the certainty of B if the risk of C is sufficiently high.
4. *Independence.* For all non-trivial probabilities, p (where "non-trivial" means that the probabilities are neither 0 nor 1), if $A \succ B$ then, for all C, $pA+(1-p)C \succ pB+(1-p)C$. Informally: if you prefer A to B then this fact is independent of whether there's some probability of instead receiving some third prospect.

A representation theorem shows that any agent who satisfies these axioms will behave as if they're maximising EU (cf. Fishburn, 1970). Given this background, the core of the EU argument relies on two premises:

EXPLOITABILITY	The above axioms can each be justified by an exploitability argument, showing that agents who violate the axiom can be offered a series of trades leading them to a guaranteed loss (such a sequence of trades is sometimes called a money pump).
AVOIDANCE	AGI will not be exploitable in this way.

Establishing EXPLOITABILITY would require introducing four exploitability arguments, one corresponding to each axiom. Meanwhile, as we'll see in §6, the argument for AVOIDANCE relies on the claim that substantial work will likely be invested into making AGIs competent, along with the claim that because exploitability undermines competence, this work will

ensure that AGI isn't exploitable. I'll consider these premises more carefully below, but for now I simply note that together they entail that AGIs will satisfy all four axioms and so, given a representation theorem, that they'll behave as if maximising EU.

To these premises, I add:

INFORMATIVENESS	Knowing that AGIs will behave as if maximising EU helps us to make practically useful predictions about what actions AGIs will undertake.

INFORMATIVENESS isn't technically necessary for the EU argument, as this argument only purported to show that AGIs could be modelled as EU maximisers, not that doing so was useful. Nevertheless, it's worth discussing this further condition because there's little point in simply having an abstract model of AGI behaviour. Instead, we want a model that can help us make useful predictions about how AGIs will behave in the real world.

Together, these three premises provide a case for modelling AGIs as EU maximisers.

An argument with roughly this shape was first gestured at by Omohundro (2008) and developed by Yudkowsky (2015), and discussion has continued since (cf. Shah, 2018; Yudkowsky, 2019; Ngo, 2019; Grace, 2021; Thornley, 2023). The above presentation represents a particularly natural and straightforward way of making such arguments precise. As a result, reflection on the EU argument provides an entry point not just for discussing this specific argument but also for gaining insight into the broader class of related arguments (and into the more general issue of how to model AGI).

So I turn now to this reflection. In the remainder of the paper, I will evaluate each of the above premises in turn.

5 EXPLOITABILITY

I start with EXPLOITABILITY, according to which each axiom from the representation theorem can be supported by an exploitability argument. Reflection on the existing literature quickly reveals that this premise is false (or at least unsupported): while exploitability arguments can be provided for completeness, transitivity, and independence, no one has provided such an argument for continuity (Gustafsson, 2022).[10]

[10] This inference from the existing literature is a little fast: this literature focuses on whether exploitability reveals irrationality, whereas my interest is not in rationality but in predicting AGIs

5.1 Continuity and Quasi-Exploitability

To clarify things here, it will be helpful to consider the closest thing we have to an exploitability argument for continuity. In particular, this argument shows that an agent who violates continuity will be willing to make a fixed payment in return for an arbitrarily small chance of some benefit. This is not an exploitability argument, because it doesn't involve a guaranteed loss. Still, we might call it a *quasi-exploitability argument*.

As context for outlining this argument, remember that an agent satisfies continuity when for all A, B, and C, if A≻B≻C then there are non-trivial p and q, such that pA+(1-p)C≻B≻qA+(1-q)C. There are various ways an agent might violate this axiom, but considering one will suffice for my purposes (for a more thorough treatment, see Gustafsson, 2022, ch. 6). I'll focus on a case where A≻B≻C but where for all p, B≻pA+(1-p)C. That is, even though A≻B, the agent will always choose B rather than risking ending up with C for the opportunity of A, however low the risk.

The quasi-exploitability argument then relies on an assumption, itself a sort of continuity requirement.[11] In particular, the argument assumes *the unidimensional continuity of preferences,* according to which if A≻B then there's some A-, such that A≻A-≻B.[12] (This claim is distinct from the continuity axiom, and so there's no circularity.)

So far then, we have the following:

(1) A≻B≻C (by assumption);

---

behaviour. Perhaps results from the existing literature won't apply in this new context, maybe because these results rely on assumptions that are true of rationality but false as claims about what we should expect from AGIs (for discussion, see Thornley, 2023).

Still, as it happens, I find it plausible that the relevant existing results also apply in the current context. While I won't explore this claim in detail, I'll discuss some relevant considerations in note 12. In any case, continuity represents the most straightforward point of failure for EXPLOITABILITY, and so that's where I'll focus my attention.

(One point of disconnect with the existing literature is worth noting: I'll restrict my attention to finite sequences of decisions, as this restriction is appropriate in the context of modelling AGI. As such, I'll set aside the cases discussed in Arntzenius et al., 2004; Bartha et al., 2014.).

[11] I draw this argument from Gustafsson, 2022, ch. 6. See also Hammond, 1998.

[12] In the current context, this assumption is a claim about the behavioural dispositions of AGIs (my interest is in what follows if AGIs fail to satisfy continuity and so what matters is whether the underlying assumptions are satisfied by AGI). Given this, it might be denied that a finite system can display the infinite richness required by the assumption.

However, concerns about the computational tractability of EU maximising arise more generally, rather than posing a particular challenge to the unidimensional continuity of preferences. So, I'll discuss this concern in more general terms in §6.3. For now, what matters is that *even if* unidimensional continuity of preferences is granted, we don't get an exploitability argument for continuity (consequently, it doesn't matter whether we accept or reject this assumption; EXPLOITABILITY is unsupported either way).

(2) B≻pA+(1-p)C for all non-trivial p (again, by assumption); and

(3) A≻A-≻B (from (1) and the unidimensional continuity of preferences).

We can then derive:

(4) A≻A-≻pA+(1-p)C for all non-trivial p (from (2), (3) and transitivity).

A quasi-exploitability argument can now be presented for continuity (or, at least, against the specific sort of continuity violation under consideration). Consider an agent who begins with the prospect pA+(1-p)C but is offered a chance to trade this for A-. They'll accept this trade, because (4) entails that A-≻pA+(1-p)C. However, because we can set p arbitrarily high, the agent will therefore be willing to accept the fixed cost of ending up with A- rather than A to avoid an arbitrarily small chance of ending up with C. To make this point more concrete, imagine that A- is the same as A, except that the agent must pay 1 penny. In that case, the agent is willing to pay 1 penny to avoid a 10% risk of C, or to avoid a 1% risk, or an 0.1% risk, or an 0.01% risk and so on *ad infinitum*. The 1 penny payment remains fixed, even as the risk avoided becomes arbitrarily small. The agent has been quasi-exploited.

Still, think what you will of this case, it doesn't involve a guaranteed loss and so is not an exploitability argument. As things stand, it follows that we're unable to offer an exploitability argument for this axiom. EXPLOITABILITY is false (or, at least, we lack grounds for thinking it's true).

5.2 Lessons

This result can be constructive as well as destructive: reflecting on the failure of the case for EXPLOITABILITY can point us towards more plausible variants on the EU argument. In particular, there are three natural responses here.[13] I lack the space to discuss these in detail, but each is worth a brief comment.

First, we might claim that not only will AGIs avoid exploitation, they'll also avoid quasi-exploitation. If so then it would suffice for a modified version of the EU argument

[13] These responses proceed most straightforwardly given exploitability arguments for the other three axioms. Gustafsson, 2022 explores such arguments in the context of discussing irrationality, but it has been argued that in the context of discussing AGI, no exploitability argument can be provided for completeness (Thornley, 2023). However, my purpose is different to Thornley's: I'm focused on modelling AGI behaviour, while Thornley is focused on whether AGI will actually decide by maximising EU. In my context, I suspect the exploitability argument for completeness succeeds. Still, I can't do justice to this issue here, so I'll simply note that even if no exploitability argument can be provided for completeness then some combination of the below responses could succeed, though modifications would be needed.

that each axiom can be supported by either an exploitability argument or a quasi-exploitability argument.

Such an approach might call for exploration, but I'm sceptical of its promise, as it's far from clear that AGIs will necessarily avoid quasi-exploitation. For example, imagine an AGI designed to promote wellbeing, which assigns disvalue to human suffering in proportion to the number of people suffering. Such a system might assign infinite disvalue to the suffering of infinitely many people, and so be willing to pay some fixed finite cost to avoid any chance whatsoever (10%, or 1%, or 0.1%, or…) of this sort of suffering. In other words, it seems plausible that an AGI of this sort might be quasi-exploitable. So it's unclear that AGI will necessarily avoid quasi-exploitability.

Further, it's not clear that being quasi-exploitable will provide any easy way for adversaries to take advantage of the AGI (and so not clear that AGI will avoid quasi-exploitability in order to avoid being taken advantage of). Of course, an adversary might attempt to make a Pascalian offer, perhaps promising infinite value in return for the one time, low payment of £100 (for a related, although finite case, see Bostrom, 2009). However, the agent will need to consider whether this is the best way to use £100 to promote infinite value, especially given the chance the adversary is lying. Plausibly, it won't be. So it's far from clear that assigning infinite value, in a way that conflicts with continuity, will allow an adversary to take advantage of an agent. The claim that AGIs will avoid quasi-exploitability would require careful argument, and absent such argument, I see no reason to accept it.

Turning to a second response to EXPLOITABILITY's failure, we might simply abandon continuity. After all, we can get sophisticated formal models of agents without requiring that continuity be satisfied (Hausner & Wendel, 1952; Hausner, 1953; Fishburn, 1971). Indeed, we can get such models even if we drop both continuity and completeness (see lemma 4.3 in McCarthy et al., 2020). By appealing to one of these models, a variant of the EU argument could be developed that appealed to a more restricted set of axioms. According to this variant argument, we should expect AGI to satisfy the more restricted set of axioms (because, as in the original argument, we should expect AGI to avoid exploitability) and hence should expect the behaviour of AGI to be described via the formal model that follows from these axioms.

In one sense this is a powerful solution, in that it makes irrelevant the question of whether continuity can be justified, whether via an exploitability argument or otherwise. On the other hand, because this approach places comparatively minimal constraints on the preferences of AGIs, it's natural to worry that it might prove uninformative. The model

might be so unconstraining that it doesn't tell us much about how AGIs will behave. So if this approach is taken, it will be necessary to provide a more detailed account of how these axiomatically-minimal models are informative (§7 will discuss some relevant considerations).

Finally, as a third response we might retain all four axioms, but not attempt to justify all via exploitability arguments. For example, we could justify completeness, transitivity, and independence in this way, and then justify continuity as a modelling assumption that allows us to provide a simple, formal model of AGI. Of course, in applying the model, it would be important to keep in mind that we'd made this assumption. Still, the presence of a simplifying assumption needn't preclude the model providing insight. Such an approach strikes me as worth taking seriously.

So EXPLOITABILITY is false, but the manner of its collapse can fruitfully point us towards various more promising alternatives to the EU argument. While my discussion of these alternatives has been brief, I hope to at least have pointed in the direction of where further reflection might bear fruit.

6 AVOIDANCE

Despite the failure of EXPLOITABILITY, it remains worth discussing AVOIDANCE. After all, as with EXPLOITABILITY, we can draw useful lessons from AVOIDANCE's failure: we can learn something about what to expect from AGI and how to model these systems. In addition, above I pointed to variants on the EU argument that might avoid the problems arising for EXPLOITABILITY. An evaluation of AVOIDANCE helps with assessing these variant arguments. So I turn to AVOIDANCE, according to which AGI won't be exploitable.

The argument for this premise can be framed in terms of optimisation power, where this is a measure of the amount of work and resources invested into making a system capable. The idea is that lots of optimisation power will be applied in the process of developing an AGI, and as a result, we should expect that AGI will be highly capable. Insofar as exploitability interferes with capability, we should therefore expect that an AGI won't be exploitable.

Spelling this out a little more, consider the process that's likely to be involved in developing AGI. First, humans will carry out a large amount of work, aimed at producing

more and more capable systems.[14] Second, there will be strong selection pressures, with the most capable systems being retained, deployed and improved, while less capable systems fall by the wayside. Some of this selection might be done using evolutionary algorithms, while other parts might arise more organically from humans choosing which systems to work on and from economic incentives channelling money into more capable systems. Finally, it's likely that AI will itself be used to make AI more competent, perhaps by generating training data or discovering effective training algorithms.[15] These three strands—human, selection pressure, and AI—will likely make systems increasingly competent. Assuming that exploitability interferes with competence, this might seem likely to push systems away from exploitable behaviour. This provides a case for AVOIDANCE.

This argument has some force: if substantial optimisation power is applied in developing AGI then this will plausibly make such systems less exploitable than they would have been given less optimisation power. Nevertheless, I doubt the argument supports anything as strong as AVOIDANCE. Indeed, I doubt that any argument supports this premise. In this section, I'll argue for this sceptical position.

I'll start by raising initial grounds for scepticism (§6.1). Then I'll argue that the benefits of immunity to exploitability are more limited than it might seem (§6.2) and the costs more substantial (§6.3). I'll conclude that the costs of full immunity likely outweigh the benefits, so we should not expect optimisation pressure to lead to full immunity to exploitability.

6.1 Initial Considerations

I'll start with some handwavey arguments. While these will be far from decisive, they're suggestive in a way that lays groundwork for further discussion.

As a first wave of the hand, note that some institutions that we might expect to be well optimised seem to be exploitable.[16]

---

[14] This will include work aimed at improving AI techniques (for example, work developing new algorithms for assigning reward to reinforcement learning systems). It will also include work aimed at making specific systems more capable, including running training processes.

[15] Early work on this issue often focused on self-modification (cf. Omohundro, 2008). Roughly, the argument was: because AI is code, self-modification is easier for AI systems than humans; AGIs will typically be incentivised to self-modify so as to avoid exploitability; so we should expect AGIs to self-modify to avoid exploitability. However, it's unclear to me whether we should actually expect such self-modification, so I focus on the argument above.

[16] I owe this point to Phil Trammell.

For example, companies operate in a ruthless commercial environment that we might expect to exert substantial optimisation pressure pushing towards sound decisionmaking. Yet many company boards make decisions by majority vote, and it's well known that majority vote can lead to violations of transitivity. Imagine Alice, Bob, and Carol on a three-person board. Alice and Bob might prefer X to Y (so, by majority vote, the board has this preference), Bob and Carol might prefer Y to Z (and so the board has this preference too), but Carol and Alice might prefer Z to X (and so the board also has this preference). Each person's preferences might satisfy transitivity, but when they decide by majority vote we can get violations of transitivity. Why then do boards decide by majority vote, given that violations of transitivity are exploitable? Presumably because such exploitability doesn't pose serious challenges in reality, or because majority vote has benefits that outweigh the costs.

Likewise, democracies operate on an international stage where there's substantial competition for economic, political and military power. Yet democratic procedures can lead to exploitability, because (to focus on just one potential mechanism) as the governing party changes, the preferences of the government change.[17] This can, in effect, lead to violations of transitivity: one government prefers X to Y, the next Y to Z, and the next Z to X. Why are some countries nevertheless democracies? One natural explanation is that this happens because, once again, the costs are either not high, or there are counterbalancing benefits, or both.

So we have some empirical evidence that successful and well-optimised institutions need not be constituted so as to avoid exploitability. While this hardly tells us anything very concrete, it might make us more sceptical that AGI will avoid exploitability.

For a wave of my other hand, it might be argued that more intelligent beings are actually more likely to be exploitable than less intelligent beings.[18] If so then the optimisation pressure designed to make AGI more intelligent might increase, rather than decrease, the chance of these systems being exploitable.

One reason to think this is empirical: it has been observed in the past that while humans violate the axioms of EU theory in various ways, many nonhuman animals at least approximately satisfy these axioms (see Stanovich, 2013, p. 3). We might conclude that in

---

[17] For discussion of relevant issues, see Riker, 1982 and Mackie, 2003.

Of course, the preferences of the government can change in a dictatorship too, either because a new dictator takes power or because the preferences of the dictator change. Still, it seems plausible that the issue will be particularly acute in democracies, where elections are frequently held, at least if the comparison is to relatively stable dictatorships. And even if this isn't so, the question still arises of why nations don't adopt some other, more stable, system of government.

[18] For discussion in the context of AGI, see Sohl-Dickstein, 2023.

biological beings, more intelligence has led to greater exploitability, and so might doubt that more optimisation pressure will necessarily mean less exploitability.

Another reason to wonder about this possibility is theoretical: a more intelligent agent will typically be better able to represent nuances of the world and develop preferences that are responsive to these nuances. While this plausibly has benefits, in terms of an improved capacity to navigate the world's complexity, it also has a cost: this complexity of preferences plausibly makes it more difficult to satisfy any axioms constraining the structure of the preferences. That is, it's plausibly harder to satisfy such axioms when one has finer-grained preferences across a finer-grained model of the world; there are more things that can go wrong with these preferences and it's more difficult to recognise when this is happening (see Stanovich, 2013; Thorstad, 2021, §5; Thorstad, Unpublished).

I don't take any of this to be decisive. Nevertheless, I do take it to put pressure on the assumption that more optimisation pressure must lead to less exploitability.

6.2 Benefits Constrained

Turning to more concrete considerations, I'll argue that the benefits of immunity to exploitability are less substantial than they might seem to be.

My first point is simple: there are no benefits to being immune to purely hypothetical exploitation (cf. Vineberg, 2022, §1.4).[19] That is, an AGI might be exploitable if their preferences violate certain axioms, but this matters only if this exploitability is likely to be… well, exploited. For this to be likely, the AGI's preferences will plausibly need to be known by many agents and some of these agents must want to take advantage of the exploitable preferences.[20] The agents will also need to be capable of carrying out the exploitation. For example, if an AGI has exploitable preferences around the possession of some rare mineral then an agent can only take advantage of this given access to the mineral. So there's no clear case that immunity to exploitability is beneficial, and even if it is, no clear case that it will be extremely beneficial rather than triflingly so.

---

[19] For discussion in the context of AGI, see Bostrom, 2012, p. 79.

[20] The agent could be exploited without the presence of any exploiter, if the agent's preferences combined with external events naturally lead the agent to make an undesirable sequence of choices because of axiom violations. However, this justifies only a limited immunity to exploitability that's likely to arise naturally. On the other hand, an adversary could potentially target any exploitable preferences, so the possibility of adversaries might seem to support the case for full immunity to exploitability and hence for AVOIDANCE.

In addition, even if an AGI is immune to exploitability then it's not actually immune to exploitability. (An apparent contradiction that needs unpacking.) Exploitability, in the sense of this paper, is a narrow notion: roughly, an agent is exploitable insofar as the formal structure of their preferences means that they'd be willing to make a series of decisions that lead to a guaranteed loss. Yet in real life, when others take advantage of us, it's not typically via this mechanism. Exploitative labour relies on economic power, or in extreme cases, threats and use of violence. Fraudsters feed us false information to manipulate us. Investors rely on a superior understanding of the stock market to make money at the expense of the less well-informed. Nations rely on the threat of military intervention to make demands on other nations. In general, we would remain exploitable by others even if our preferences satisfied the axioms of EU theory. The same is true of AGI. This suggests that if an agent is eager to exploit an AGI then there's no safety in axiom satisfaction; instead, the agent can simply choose some other way of taking advantage of the AGI. Immunity to exploitability, in the narrow sense of this paper, doesn't serve the protective function it might seem to, or at least it doesn't do so robustly.

None of this reduces the benefits of immunity to exploitability (in the narrow sense) to zero, as long as there's some practical possibility of exploitation occurring and some chance that the would-be-exploiter won't use some alternative means of exploitation.[21] Still, in combination, these considerations show that the benefits of immunity are less substantial than it might seem.

6.3 Costs Accrued

While the benefits are more meagre than might have been expected, the costs are more substantial, for three reasons.

First, the optimisation power applied to ensure that an AGI's preferences satisfy the axioms could have instead been used to achieve something else. For example, I noted above that an agent can be taken advantage of by those with a superior understanding of the stock market. So rather than ensuring greater concordance with the axioms, optimisation pressure might instead be fruitfully used to increase an AGI's capacity to reason about financial

[21] One question I haven't explored is whether exploitation of AGI by AGI is likely to be different to exploitation of humans by humans. For example, the possibility of an adversary gaining access to the AGI's source code might make certain forms of exploitation more likely, and so might change the benefits of immunity to exploitation.

matters. Ensuring that an AGI's preferences satisfy the EU axioms comes at an opportunity cost.[22]

Second, avoiding exploitation by explicitly deciding using the full formal apparatus of EU theory is computationally intractable, at least if an agent relies on sufficiently fine-grained ways of characterising possible states of the world (see van Rooij et al., 2018, p. 494).[23] Further, it has been argued that the same is true of agents that merely decide *as if* maximising EU. After all, if an agent would always decide as if maximising EU then the agent must, in some fashion, have solved the intractability problem (because their dispositions represent a solution to the problem). Given that solving the problem is intractable, it follows that we should not expect AGIs to act as if maximising EU (see Bossaerts et al., 2018; van Rooij et al., 2018).

This might seem to force us to reject AVOIDANCE. However, this would be an overreaction. For a start, it's computationally tractable to decide by maximising EU as long as one has a sufficiently coarse-grained representations of world states and acts. This in itself comes at a cost (in terms of being able to represent complexities in the world), but it does reveal that the issue is more complex than outright intractability. Further, even insofar as it's computationally intractable to decide as if maximising EU, there may be computationally tractable processes that approximate an agent deciding in this fashion. EU theory will provide a good (if imperfect) model of a system that utilises such approximations. Consequently, the EU argument could be reframed as the claim not that AGIs will behave *as if* maximising EU, but instead as the claim that they will approximate such behaviour. In this case, computational considerations aren't decisive. Still, it remains the case that there are computational costs to satisfying the axioms (or approximating doing so to a close degree). This cost is real, even if it's not decisive in isolation.

To get to a third cost, consider a case where an AGI is applying the optimisation pressure. Concretely, we might imagine this involves the AGI making changes to the algorithms or data used to train AI systems, where this new training process could then be used to train a successor system (Ngo, 2020, pp. 6–8). From the AGI's perspective, there may be additional costs to applying optimisation power in a way that ensures that the successor satisfies the axioms: it might lead the successor system to invest resources into achieving outcomes that the AGI doesn't care about.

---

[22] This opportunity cost might fall away given sufficient optimisation pressure, as it will then be possible to optimise in all of the desired ways. However, it might be doubted that we'll see this idealised level of optimisation pressure.

[23] See Bostrom, 2012, p. 79 for discussion in the context of AI.

For example, assume that an AGI violates transitivity as a result of having cyclic strict preferences, such that $A \succ B \succ C \succ A$. Ensuring that the successor system satisfies transitivity will require that this successor lacks one of the strict preferences that the AGI possesses. For illustration, let's imagine the successor system has a preference for A over C (for a more general discussion, see Williamson, 2022, pp. 181–182). In this case, the successor will be willing to make a payment in order to receive A rather than C. Yet from the AGI's perspective, such a payment is not just pointless but counterproductive.[24] After all, the AGI prefers C to A. So from the AGI's perspective, there will be a cost to its successor satisfying transitivity; applying optimisation pressure in this way is not straightforwardly beneficial.

So there are costs to immunity to exploitability and the benefits are weaker than it might seem. Plausibly, I suggest, the costs outweigh the benefits, and so we should not expect AGI to satisfy the axioms. In combination with the considerations from §6.1, this gives us grounds to reject AVOIDANCE.

6.4 Lessons

One response to the above discussion would be to deny that the costs truly outweigh the benefits. However, I'll assume this claim is right and ask how those who agree might respond to the failure of AVOIDANCE.

I've already pointed to a partial resolution above. AVOIDANCE was framed in absolute terms: AGI *will* avoid exploitation. However, this claim could be weakened by moving to talk of approximations. The claim then becomes that an AGI will *approximate* an agent that avoids exploitability, within the limits allowed by computational constraints.

This claim could be further weakened. After all, the computational tractability of deciding by maximising EU can be improved still further by narrowing the range of cases where the agent behaves in this way (van Rooij et al., 2018, §3). Perhaps, for example, AGIs will behave as if maximising EU only in those situations most likely to be encountered, rather than in all possible situations.[25] Or perhaps AGIs will decide as if maximising EU in just those cases where they're most likely to be exploited. Given that AGI will operate in a world of humans, this might involve immunity to forms of exploitability that humans are

[24] Here, I'm assuming the AGI cares about the various outcomes in an agent-neutral fashion. If this isn't so then my argument here won't apply.

[25] In the current machine learning paradigm, the optimisation power used to develop AGI is applied in a training environment. Consequently, we might expect the system to approximate EU maximisers in environments similar to the training environment. See Dai, 2019.

capable of taking advantage of. If so, AGI might behave as if maximising EU in most interactions with humans (see Yudkowsky, 2015). This would bring many of the benefits of general immunity to exploitability with fewer costs.

Finally, we could acknowledge that the situation is complicated, with hard to assess costs and benefits in play. As a result, we might step back from a claim about what AGI will be like and might instead settle for a claim that considerations of exploitability are at least likely to push AGI in a certain direction. On this view, consideration of exploitability suggests that AGI are likely to more closely approximate EU maximisers than they would otherwise have done (without this necessarily implying that the approximation will be particularly close). Of course, this weakens what can be concluded from the argument, but it might nevertheless feed into our broader reflections on the likely nature of AGI.

Putting all of this together, a weaker form of AVOIDANCE might state that AGI will be pushed in the direction of more closely approximating agents that are immune to exploitation, especially in cases that are particularly likely to be encountered and even more so in cases where exploitation is a serious risk.

7 INFORMATIVENESS

I turn now to INFORMATIVENESS, according to which we would learn something useful about AGI if we knew that these systems would act as if maximising EU. We have immediate grounds to reject this claim: knowing that an agent will decide as if maximising EU tells us nothing at all about this agent's behaviour, because any and all behaviour can be modelled as EU maximising behaviour.

7.1 The Triviality of EU

The basic point is that EU theory gives us two unconstrained variables that we can use to explain an agent's behaviour: the probabilities and utilities that are used in calculating the EUs. And any behaviour can be explained as if it were EU maximising behaviour simply by stipulating appropriate values for these probabilities and utilities.

Even apparently clear violations of the axioms can be explained in terms of EU maximising. For example, consider an agent who starts with an apple, trades this for a banana for a 1 penny cost, trades this for a carrot for 1 penny, and then trades this back for an apple for 1 penny. This agent appears to violate transitivity, preferring carrots to bananas

to apples to carrots. However, the agent's behaviour can instead be explained as indicating that they have time-sensitive utilities. They assign a higher utility to "a banana at time 1" to "an apple at time 1", a higher utility to "a carrot at time 2" to "a banana at time 2", and a higher utility to "an apple at time 3" to "a carrot at time 3". By describing outcomes and stipulating utilities carefully, we avoid any violation of transitivity and this agent can be modelled as an EU maximiser.[26]

More generally, for any course of action that an agent carries out, we can simply specify that the agent's utility function assigns positive utility to carrying out each of these actions and assigns a utility of 0 to everything else, including all other actions that could have been undertaken. Given this specification of the utility function, it trivially follows that the agent is acting so as to maximise EU. Consequently, however an agent behaves, they can be interpreted as if they're maximising EU.

Given that all behaviour is compatible with an agent behaving as if maximising EU, learning that an AGI will behave in this way tells us nothing about how AGI will behave. INFORMATIVENESS is false in the most extreme way possible.

7.2 Lessons

The failure of INFORMATIVENESS resulted from the fact that the utility function is a free variable in EU theory: as long as we can freely specify the agent's utility function, all behaviour can be modelled as EU maximising. It follows that if an EU model is to be informative then we must constrain the utility functions that can be attributed to AGIs.[27] With sufficient constraints, it will become informative to learn that AGIs will act as if maximising EU by the lights of some utility function of the specified sort. This need for utility-function constraint is the core lesson to be taken from the failure of INFORMATIVENESS.

Of course, it would be no help to arbitrarily constrain the utility functions. Instead, it'll need to be plausible that the constraints imposed accurately characterise AGI. Under

[26] Cf. Pettit, 1991, p. 163; Broome, 1993; Hodgson, 2012. In the context of AGI, see Shah, 2018; Drexler, 2019, §6.4.

[27] In fact, if the probability function is unconstrained then, even if we constrain the utility function, it will remain uninformative to know that some agent can be modelled as if maximising EU. So it'll also be necessary to constrain what probability functions AGIs will have (for example, we might argue that AGIs will be fairly well calibrated and their probabilities will approximately track the evidential probabilities). In any case, I'll focus on the utility function in my discussion.

these circumstances, learning about the implications of the constrained utility functions tells us something about AGI.

While this approach can rescue the informativeness of the EU argument, it comes at a cost. Previously, it might have been hoped that the EU argument would allow us to make predictions about AGI from minimal assumptions. However, the above solution requires us to build substantive knowledge of AGI into our model from the outset. And, as noted earlier, it's hard to accurately speculate about hypothetical future AI systems, so the need for substantive knowledge is a serious cost.

Still, there are approaches that might allow us to gain insight into the utility functions that are likely to characterise AGI. One possibility would be to rely on the assumption that AGI will be developed using recognisable descendants of current techniques. If so then work exploring current techniques might provide insight into the likely shape of AGI. Another possibility would be to appeal to more abstract considerations. For example, one could argue that AGIs will plausibly have (or be best modelled as having) relatively simple utility functions, perhaps because such functions are computationally tractable or because successful agents are generally likely to have a simplicity bias. Either of these approaches might allow us to identify plausible constraints on the utility functions of AGIs, and so might allow us to develop informative variants on the EU argument.[28]

So, as with the other premises, reflection on the failure of INFORMATIVENESS points towards more nuanced and sophisticated variants on the EU argument.

8 Modelling AGI

The EU argument fails. EXPLOITABILITY, AVOIDANCE and INFORMATIVENESS are all false.

Nevertheless, from the failure of these premises, we learn something about what it would take to develop a more nuanced and plausible variant on the EU argument. Such an

[28] This gets us to informativeness only in that the argument rules out certain behaviours. More would need to be said to show that we learn something *important*. One possibility would be to argue that AGIs are likely to have (or behave as if they have) broadly-scoped utility functions, where these encode goals that apply across long time scales and large spatial areas, and which aren't easily satiated. For example, consider a utility function that assigns 1 utility for each chess game won. This makes no reference to where or when the games are won and places no limits on how many games the system should attempt to win. It's therefore broadly-scoped.

It might then be argued that broadly-scoped utility functions will likely give rise to instrumentally convergent subgoals of the sort discussed in Bostrom, 2012. These subgoals include resource acquisition and survival, and so it would follow that AGIs would likely seek to acquire resources and survive. This result would be not just informative but also important. See Ngo & Bales, Forthcoming, §5.

argument is likely to either rely on a smaller set of axioms or to justify some axioms on grounds other than exploitability. It's likely to draw a weaker conclusion, suggesting that AGI will be pushed in the direction of approximating an EU maximiser, rather than outright act as if maximising EU. And it will rely on making, and defending, substantive claims about the likely shape of utility functions that AGI will be best modelled as possessing. While I've hardly fully developed such an argument here, I hope to have provided the foundations from which such development could proceed.

However, the results of this paper also point to more general lessons.

For a start, once we develop substantive knowledge of the utility functions that plausibly best model AGI (per §7.2), we might wonder whether this substantive knowledge provides greater insight than the EU model itself. Perhaps what matters isn't whether AGI is well characterised via some very specific mathematical formalism. Instead, perhaps what matters is whether there's some broad sense in which AGIs are likely to be goal-directed and instrumentally rational, and if so what sorts of goals AGIs are likely to pursue. So instead of attempting to develop more sophisticated variants on the EU argument, we might look to develop a more sophisticated understanding of this broader picture about AGI goals.[29]

I think there's something to this suggestion but also suspect it's too binary, insofar as it suggests we must make a choice between EU models and broader models of goal-driven agents. Instead, I think one lesson we might draw from the failure of the EU argument is that we shouldn't be looking for one single approach that provides the true model of AGI. It's likely to be more fruitful to use a variety of models, each of which provide some evidence and place some constraints on our expectations. We might use EU theory, alongside more informal notions of goal-directedness. We might also use game theory to model interaction between agents, including humans, institutions, and AGIs. And we might use the predictive processing model, familiar from cognitive science, to model not humans but AGIs (cf. Ratoff, 2021). In combination, these models, and others besides, might allow us to build a broader picture.

Of course, even in combination, these models are likely to prove limited. After all, in modelling AGI we are speculating about a future technology that is, in many ways, unprecedented. In such a context, we should expect models to be simplifications, to miss important considerations, and possibly to be outright misleading. Overall, we should think of such models as small islands of evidence in a vast sea of uncertainty. This might feel

[29] See Ngo & Bales, Forthcoming and Unknown Author, Unknown Date (I suspect the author is Eliezer Yudkowsky).

somewhat underwhelming as the final words in a paper on modelling AGI. Still, I believe we learn more from recognising our limitations than from imposing false certainty.

Bibliography

Arntzenius, F., Elga, A., & Hawthorne, J. (2004). Bayesianism, Infinite Decisions, and Binding. *Mind*, *113*(450), 251–283.

Bartha, P., Barker, J., & Hájek, A. (2014). Satan, Saint Peter and Saint Petersburg: Decision theory and discontinuity at infinity. *Synthese*, *191*(4), 629–660.

Bossaerts, P., Yadav, N., & Murawski, C. (2018). Uncertainty and computational complexity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1766).

Bostrom, N. (2009). Pascal's mugging. *Analysis*, *69*(3).

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, *22*(2), 71–85.

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

Broome, J. (1993). Can a Humean be moderate? In C. W. Morris & R. G. Frey (Eds.), *Value, Welfare, and Morality* (pp. 51–73). Cambridge University Press; Cambridge Core.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. arXiv. https://arxiv.org/abs/1802.07228

Carlsmith, J. (Forthcoming). Existential Risk From Powerseeking AI. In D. Thorstad, J. Barrett, & H. Greaves (Eds.), *Essays on Longtermism*. Oxford University Press.

Clarke, C. (2016). Preferences and Positivist Methodology in Economics. *Philosophy of Science*, *83*(2), 192–212.

Cotra, A. (2020). *Forecasting TAI with Biological Anchors (Draft Report)*. Open Philanthropy.

https://drive.google.com/drive/u/1/folders/15ArhEPZSTYU8f012bs6ehPS6-xmhtBPP

Dafoe, A. (2018). *AI Governance: A Research Agenda* (v1.0). Centre for the Governance of AI.

Dai, W. (2019). Three Ways That 'Sufficiently Optimized Agents Appear Coherent' Can Be False. *AI Alignment Forum*. https://www.alignmentforum.org/posts/4K52SS7fm9mp5rMdX/three-ways-that-sufficiently-optimized-agents-appear

Drexler, K. E. (2019). *Reframing Superintelligence: Comprehensive AI Services as General Intelligence* (Technical Report #2019-1). Future of Humanity Institute, University of Oxford.

Fishburn, P. (1970). *Utility Theory for Decision Making*. Wiley.

Fishburn, P. C. (1971). A Study of Lexicographic Expected Utility. *Management Science*, *17*(11), 672–678.

Garrabrant, S. (2022). Comment on 'Why the Focus on Expected Utility Maximisers?' *Less Wrong*. https://www.lesswrong.com/posts/XYDsYSbBjqgPAgcoQ/why-the-focus-on-expected-utility-maximisers?commentId=a5tn6B8iKdta6zGFu

Grace, K. (2021). Coherence Arguments Imply a Force for Goal-Directed Behavior. *AI Alignment Forum*. https://www.alignmentforum.org/posts/DkcdXsP56g9kXyBdq/coherence-arguments-imply-a-force-for-goal-directed-behavior

Gustafsson, J. E. (2022). *Money-Pump Arguments*. Cambridge University Press; Cambridge Core.

Hammond, P. (1998). Objective Expected Utility: A Consequentialist Perspective. In S. Barberà, P. Hammond, & C. Seidl (Eds.), *Handbook of Utility Theory Volume 1: Principles* (pp. 143–211). Kluwer.

Hausner, M. (1953). *Multidimensional Utility* (No. 604151). Rand Corporation.

Hausner, M., & Wendel, J. G. (1952). Ordered Vector Spaces. *Proceedings of the American*

Mathematical Society, *3*(6), 977–982.

Hodgson, G. M. (2012). On the Limits of Rational Choice Theory. *Economic Thought*, *1*(1).

Korinek, A., & Juelfs, M. (Forthcoming). Preparing for the (Non-Existent?) Future of Work. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford Handbook of AI Governance* (p. 0). Oxford University Press.

Mackie, G. (2003). *Democracy Defended*. Cambridge University Press.

McCarthy, D., Mikkola, K., & Thomas, T. (2020). Utilitarianism with and without expected utility. *Journal of Mathematical Economics*, *87*, 77–113.

Ngo, R. (2019). Coherent Behaviour in the Real World is an Incoherent Concept. *AI Alignment Forum*. https://www.alignmentforum.org/posts/vphFJzK3mWA4PJKAg/coherent-behaviour-in-the-real-world-is-an-incoherent

Ngo, R. (2020). AGI Safety from First Principles. *AI Alignment Forum*. https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ

Ngo, R., & Bales, A. (Forthcoming). Deceit and Power: Machine Learning and Misalignment. In D. Thorstad, J. Barrett, & H. Greaves (Eds.), *Essays on Longtermism*. Oxford University Press.

Omohundro, S. M. (2007). *The Nature of Self-Improving Artificial Intelligence*. https://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf

Omohundro, S. M. (2008). The Basic AI Drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*. IOS Press.

Peterson, M. (2017). *An Introduction to Decision Theory* (2nd Edition). Cambridge University Press.

Pettit, P. (1991). Decision theory and folk psychology. In M. Bacharach & S. Hurley (Eds.), *Essays in the Foundations of Decision Theory* (pp. 147–175). Blackwell.

Ratoff, W. (2021). Can the predictive processing model of the mind ameliorate the value-alignment problem? *Ethics and Information Technology*, *23*(4), 739–750.

Riker, W. H. (1982). *Liberalism Against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice*. Waveland Press.

Roser, M., Herre, B., & Hasell, J. (2013). Nuclear Weapons. *Our World in Data*.

Shah, R. (2018). Coherence Arguments Do Not Entail Goal-Directed Behavior. *AI Alignment Forum*. https://www.alignmentforum.org/posts/NxF5G6CJiof6cemTw/coherence-arguments-do-not-entail-goal-directed-behavior

Sohl-Dickstein, J. (2023). The Hot Mess Theory of AI Misalignment: More Intelligent Agents Behave Less Coherently. *Jascha's Blog*. https://sohl-dickstein.github.io/2023/03/09/coherence.html

Stanovich, K. E. (2013). Why humans are (sometimes) less rational than other animals: Cognitive complexity and the axioms of rational choice. *Thinking & Reasoning*, *19*(1), 1–26.

Stein-Perlman, Z., Weinstein-Raun, B., & Grace, K. (2022). *2022 Expert Survey on Progress in AI*. AI Impacts. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai

Thornley, E. (2023). There Are No Coherence Theorems. *The Effective Altruism Forum*. https://forum.effectivealtruism.org/posts/FoRyordtA7LDoEhd7/there-are-no-coherence-theorems

Thorstad, D. (2021). The accuracy-coherence tradeoff in cognition. *The British Journal for the Philosophy of Science*.

Thorstad, D. (Unpublished). *The Complexity-Coherence Tradeoff in Cognition*.

Trammell, P., & Korinek, A. (2020). *Economic Growth Under Transformative AI* (GPI Working

Paper No. 8–2020). Global Priorities Institute.

Unknown Author. (Unknown Date). Consequentialist Cognition. *Arbital*. https://arbital.com/p/consequentialist/

van Rooij, I., Wright, C. D., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese*, *195*(2), 491–510.

Vineberg, S. (2022). Dutch Book Arguments. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/dutch-book/

Williamson, T. (2022). *Dominance in Decision Theory* [Australian National University]. https://openresearch-repository.anu.edu.au/bitstream/1885/250952/1/Dominance%20in%20Decision%20Theory.pdf

Yudkowsky, E. (2015). Sufficiently Optimized Agents Appear Coherent. *Arbital*. https://arbital.com/p/optimized_agent_appears_coherent/

Yudkowsky, E. (2019). Coherent Decisions Imply Consistent Utilities. *Less Wrong*.

Zhang, B., Dreksler, N., Anderljung, M., Kahn, L., Giattino, C., Dafoe, A., & Horowitz, M. C. (2022). *Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers*. arXiv. https://arxiv.org/abs/2206.04132