Robustness to fundamental uncertainty in AGI alignment

G Gordon Worley III[1]

[1]Phenomenological AI Safety Research Institute

July 27, 2018

Abstract

The AGI alignment problem has a bimodal distribution of outcomes with most outcomes clustering around the poles of total success and existential, catastrophic failure. Consequently, attempts to solve AGI alignment should, all else equal, prefer false negatives (ignoring research programs that would have been successful) to false positives (pursuing research programs that will unexpectedly fail). Thus, we propose adopting a policy of responding to points of metaphysical and practical uncertainty associated with the alignment problem by limiting and choosing necessary assumptions to reduce the risk false positives. Herein we explore in detail some of the relevant points of uncertainty that AGI alignment research hinges on and consider how to reduce false positives in response to them.

Introduction

The development of artificial intelligence (AI) is making possible many technological advancements and improvements to wellbeing but also exposes humans to considerable risks (Brundage et al., 2018). Key among these risks are the existential risks associated with artificial general intelligence or AGI (Turchin and Denkenberger, 2018). In particular, it is likely that the development of superintelligent AGI will create either existential catastrophe or massive benefits for humanity with little room for mildly bad, mildly good, or neutral outcomes (Bostrom, 2014). This bimodal distribution of outcomes where one of the outcomes is extremely undesirable implies that we are better off if we focus more on avoiding negative outcomes and less on achieving positive outcomes (Bostrom, 2003). Thus, assuming AGI will be built eventually, we can increase the expected value of AGI by working on interventions that reduce the chance of existential catastrophe, and we call the field that focuses on these interventions AI safety (Yampolskiy, 2012).

Central among interventions that address AI safety is alignment, the problem of how to build AGI that is aligned with human interests (Soares and Fallenstein, 2017). The alignment problem is

complex and likely requires solving many open problems in mathematics, economics, computer science, and philosophy (Yudkowsky, 2016). This is unfortunate from the perspective of increasing the expected value of AGI by reducing existential risks because it introduces many opportunities for researchers to make mistakes. If those mistakes are false negatives, like believing an alignment scheme won't work when it would, then there is a comparatively small loss of value from failing to develop safe AGI as soon as possible, but if those mistakes are false positives, like believing an alignment scheme will work when it won't, then there is an astronomically large loss of value from accidentally developing unsafe AGI. This implies that, all else equal, we are better off preferring false negatives to false positives when working on alignment.

This preference suggests several AGI research and development policies. One previously explored by Yudkowsky is the idea of security mindset, borrowing an idea of the same name from computer security researcher Bruce Schneier, which we might summarize as the expectation of operating in an adversarial environment that is actively trying to produce unsafe AGI (Yudkowsky, 2017b). Another, which we will explore here, might be called robustness to fundamental uncertainty, the idea that we can reduce false positives by choosing necessary assumptions when reasoning about alignment such that the chosen assumptions minimize the risk of false positives among the space of alternative assumptions.

Reducing false positive risk

If we wish to reduce false positives in AI alignment research, viz. reduce the chance of accidentally producing unaligned AGI when we thought we would have produced aligned AGI, we can work to reduce the overall rate of errors or we can tradeoff producing more false negatives against producing fewer false positives (Neyman and Pearson, 1933). The former is hard to control since it largely depends on the base error rates of individual researchers, but the the tradeoff between false negatives and false positives may be adjusted by taking actions that prefer one to the other. For example, to decrease false negatives and increase false positives researchers might try more ideas they think less likely to work and vice versa to increase false negatives and decrease false positives. The latter is the sort of approach to reducing false positives one might get in part by application of Yudkowsky's ideas around security mindset.

Another approach to reducing the rate of false positives by trading off against false negatives is to apply the principle of parsimony (also known as Occam's razor) to decrease the probability of any particular attempt at producing aligned AGI failing by reducing the number of variables multiplied together in calculating the probability of success (Pearl, 2000). For example, suppose we have two alignment methods, A and B, and the probability of A producing aligned AGI depends on the outcome of 5 variables and B depends on the outcome of 4, then all else equal we should expect B to be more likely than A to produce aligned AGI. On the one hand this suggests the

adoption of the well-known engineering principle that simpler systems are less likely to fail, and on the other suggests we can reduce false positives by making a deep commitment to reducing the number of variables, especially implicit variables that are often ignored, when building aligned AGI (Billington and Allan, 1992).

We can reduce the number of variables by eliminating unnecessary variables that take the form of assumptions—unreliable beliefs—in our reasoning. Consider, if we had a plan to build aligned AGI that depended on variables X, Y, and Z, but suppose X could be deduced from a simpler and more likely variable W, then we should replace X with W in our reasoning. Or if Y and Z could both be deduced from another variable V, then Y and Z should be replaced by V. Such steps should obviously be taken, but they are of limited effectiveness because they do not necessarily eliminate all variables. One cause of our need to reason with irreducible variables is computational because complete deductive proofs cannot always be found within reasonable time bounds (Leike and Hutter, 2015). The other, more pernicious cause of this need is the problem of epistemic circularity.

Briefly stated, epistemic circularity is the problem that nothing can be reliably known without first knowing something reliably (Alston, 1986). There have been numerous attempts to solve epistemic circularity since it was first identified by Pyrrhonian skeptics in ancient Greece through the problem of the criterion, and it remains a key problem in epistemology today because it likely has no complete solution (Lammenranta, 2003). Instead we are left to adopt the least-bad pragmatic solution, known as particularism, by choosing to assume unreliable entitlements to some knowledge and then reasoning as if we knew those facts reliably (Chisholm, 1973), (Alston, 1993), (Alston, 2005). This does not fundamentally address the skepticism that epistemic circularity implies, but it does allow us to contain our skepticism to only a few philosophical assumptions—called hinge propositions by Wittgenstein, analogous to axioms in formal logics, and related to approximation of the universal prior in Solomonoff induction—to get on with reasoning in spite of epistemic circularity (Talvinen, 2009).

Adopting particularism still leaves us with the problem of choosing the specific assumptions upon which we will build our reasoning about how to build aligned AGI. Since these choices cannot be made reliably they stand as "free" variable in our reasoning that we must choose using criteria other than likelihood of being true since that likelihood cannot be adequately assessed. This need to choose gives us an opening, though, to choose such that the risk of false positives is reduced by considering the relative likelihood of false positives given different choices of assumptions. We will now review some of these philosophical assumptions, why they are necessary, and evaluate their expected impacts on the likelihood of false positives.

Necessary assumptions

We have identified at least two problems, both metaphysical and pragmatic, that presently require making assumptions when reasoning about AGI alignment: meta-ethical uncertainty and uncertainty about mental phenomena . For each problem we consider why making an assumption is necessary and give arguments in favor of and against particular assumptions in terms of their impact on false positives and likelihood of success.

Meta-ethical uncertainty

AGI alignment is typically phrased in terms of aligning AGI with human interests, but this hides some of the complexity of the problem behind determining what "human interests" are. Taking "interests" as a synonym for the cluster of concepts we also call "values" and "preferences", we can begin to make some progress by treating alignment as at least partially the problem of teaching AGI to learn human values (Soares, 2016), (Arnold *et al.*, 2017). Unfortunately, what constitutes human values is currently unclear since humans may not be aware of the extent of their own values and may not hold reflexively consistent or rational values (Scanlon, 2003), (Tversky, 1969). This creates a problem in terms of alignment because the values of individual humans, let alone the combined values of humanity, contradict each other, so in order for an AGI to align its behavior with human values it must have some way of resolving those conflicts (Yudkowsky, 2004). Regardless of the value conflict resolution mechanism used, having a way to resolve value conflicts amounts to making a normative decision about values, and doing so means alignment asks us to tackle the ethical issue of normative uncertainty (MacAskill, 2014).

Normative uncertainty is a symptom of deeper uncertainty in ethics caused by meta-ethical uncertainty about the existence of moral facts because the problem of the criterion prevents us from reliably knowing whether or not moral facts exist, let alone what moral propositions are true if moral facts exist (Chisholm, 1982). Meta-ethical uncertainty forces us to speculate about moral facts because knowledge about moral facts, even if it is the knowledge that moral facts do not exist, is necessary to ground ethical reasoning (Zimmerman, 2010). Thus in order to reason about alignment we must consider what hinge proposition to adopt about the existence of moral facts in order to be able to design AGI that can behave in a manner aligned with conflicting human values. The standard positions regarding the existence of moral facts are realism, anti-realism, and skepticism being respectively for, against, and uncertain about their existence, so we'll consider their effects on false positives in AGI alignment in turn.

If we suppose realism, then we could build aligned AGI on the presupposition that it could at least discover moral facts even if no moral facts were specified in advance and then use knowledge of these facts to resolve conflicts in human values. Now suppose this assumption is false and that moral facts do not exist or even if they do exist they cannot be known, then our moral-facts-

assuming AGI would either never discover any moral facts to guide its behavior when human values are in conflict or would assume arbitrary moral propositions to be facts that would not be sure to produce resolutions to value conflicts that humans would want. Such an AGI might still achieve de facto alignment with human values if it adopted moral propositions it believed to be facts that allowed it to converge on a functionally equivalent solution to the one an AGI that had correctly assumed no knowledge of moral facts would have used, but lacking an argument suggesting such an AGI would be less likely to result in false positives than a simpler AGI that started out not assuming knowledge of moral facts this seems a strictly more risky approach.

If we suppose anti-realism, then we must build aligned AGI to reason about conflicts in human values in the absence of any normative assumptions grounded in moral facts. Now suppose this assumption is false and moral facts do exist, then our moral-facts-denying AGI would resolve value conflicts in a way not based on moral facts and would fail to act in a way that fully satisfied human preferences for ethical behavior. Such an AGI might still achieve de facto alignment with human values if it adopted conflict resolution mechanisms that were functionally equivalent to those it would have adopted if it had known moral facts, and it may be able to do this because the unacknowledged moral facts would impact such an AGI through their influence on human values. Although less efficient than acting on moral facts directly, this suggests an anti-realist AGI could still stand a chance of aligning itself based on moral facts as revealed by the human values being aligned with.

If we suppose skepticism, then we must build aligned AGI in the absence of any knowledge of moral facts, which is similar to the anti-realist case, but different in that there we assume moral facts do not exist where here we remain open to the possibility that moral facts may exist even if we don't or can't know them. Now suppose this assumption is false and we can know about the existence of moral facts such that we could decide in favor of realism or anti-realism, then the AGI may fail to acknowledge and use this knowledge to make more informed decision about value conflict resolution, however we could reasonably expect this to be mitigated because skepticism, unlike realism and anti-realism, need not assume a strong metaphysical claim and only instead claim that perfect knowledge is not possible, thus the skeptical AGI could switch to believing realism or anti-realism with high credence and acting on that belief. This still leaves open the practical question of how a skeptical AGI would address value conflicts, but it could reasonably either choose norms the same way an anti-realist AGI would until it learned more, or it could adopt moral particularism to ground its norms on moral assumptions.

The result of this analysis is that it is probably best to adopt moral skepticism to reduce the risk of false positives when building aligned AGI. Although it may be a much less direct route to aligned AGI than assuming realism would be if realism were true and a slightly less direct route than assuming anti-realism if anti-realism were true, skepticism allows AGI research to avoid committing to a hinge proposition that will much raise the risk of false positives. This unfortunately makes solving alignment harder because it eliminates the opportunity to ground choices about value con-

flict resolution norms in thoroughly grounded moral reasoning, but it would still be possible to make progress by adopting norm particularism (after the style of Dancy's ethical particularism) despite not being able to make truth claims about which norms are best (Dancy, 1983). A direction of future research could include evaluating norms that might be adopted in the same way we are here evaluating hinge propositions to minimize the likelihood of false positives.

Uncertainty about mental phenomena

Since AGI alignment is necessarily alignment of AGI, alignment schemes can depend on the dispositions of AGI, and one disposition AGI has is to subjective experience and mental phenomena (Adeofe, 1997), (Nagel, 1974). Whether or not we expect AGI to realize this disposition matters because it influences the types of alignment schemes that can be considered since an AGI without a mental aspect can only be influenced by modifying its algorithms and manipulating its behavior whereas an AGI with a mind can be influenced by engaging with its perceptions and understanding of the world (Dreyfus, 1978). In other words we might say mindless AGI can be aligned only by algorithmic and behavioral methods whereas mindful AGI can also be aligned by philosophical methods that work on its epistemology, ontology, and axiology (Brentano, 1995). It's unclear what we should expect about the mentality of future AGI, though, because we are presently uncertain abut mental phenomena in general (cf. the work of Chalmers and Searle for modern, popular, and opposing views on the topic), so we are forced to speculate about mental phenomena in AGI when we reason about alignment (Chalmers, 1996), (Searle, 1984).

Note, though, that this uncertainty may not be fundamental (Dennett, 1991). For example, if materialist or functionalist attempts to explain mental phenomena prove adequate, perhaps because they lead to the development of conscious AGI, then we may agree on what mental phenomena are and how they work (Oizumi *et al.*, 2014). If they don't, though, we'll likely be left with metaphysical uncertainty around mental phenomena that's rooted in the epistemic limitations of perception (Hussrl, 2014). Regardless of how uncertainty about mental phenomena might later be resolved, it currently creates a need for pragmatically making assumptions about it in our reasoning about alignment. In particular we want to know whether or not we should design alignment schemes that assume a mind, even if we expect mental phenomena to be reducible to other phenomena. Given that we remain uncertain and cannot dismiss the possibility of mindful AGI, what we decide depends on how likely alignment schemes are to succeed and avoid false positives conditional on AGI having the capacity for mental phenomena. The choice is then between whether we design alignment schemes that work without reference to mind or whether they engage with it.

If we suppose AGI do not have minds, whether because we believe they have none, are inaccessible to us, or not causally relevant to alignment, then alignment schemes can only address the algorithms and behavior of AGI. Now suppose this assumption is false and AGI do have minds, then our alignment schemes that work only on algorithms and behavior would be expected to con-

tinue to work since they function without regard to the mental phenomena of AGI, making the minds of AGI irrelevant to alignment. This suggests there is little risk of false positives from supposing AGI do not have minds.

If we suppose AGI do have minds, then alignment schemes can also use philosophical methods to address the values, goals, models, and behaviors of AGI. Such schemes would likely take the form of ensuring that updates to an AGI's ontology and axiology converge on and maintain alignment with human interests (de Blanc, 2011), (Armstrong, 2015). Now suppose this assumption is false and AGI do not have minds, then our alignment schemes that employ philosophical methods will likely fail because they are attempting to address mechanisms of action not present in AGI. This suggests there is a risk of false positives from supposing AGI have minds proportionate with the likelihood that we do not build mindful AGI.

From this analysis it seems we should suppose mindless AGI when designing alignment schemes so as to reduce the risk of false positives, but note that it does not consider the likelihood of success at aligning AGI using only algorithmic and behavioral methods. That is, all else may not be equal between these two assumptions such that the one with the lower risk of false positives might not be the better choice if we have additional information that leads us to believe that alignment of mindful AGI is much more likely to succeed than the alignment of mindless AGI, and it appears that we have such information in the form of Goodhart's curse and the failure of good old-fashioned AI (GOFAI).

Goodhart's curse says that when optimizing for the measure of a value the optimization process will implicitly maximize divergence of the measure from the value (Yudkowsky, 2017a). This is an observation that follows from the combination of Goodhart's law and the optimizer's curse (Goodhart, 1984), (Smith and Winkler, 2006). The tendency of measure and value to diverge under optimization results in a phenomenon known as "Goodharting", and it takes myriad forms that affect alignment (Manheim and Garrabrant, 2018). In particular Goodharting poses a problem for behavioral alignment schemes because to optimize behavior it is necessarily to measure behavior and then optimize on that measure. Consequently it appears behavioral methods are unlikely to be capable of producing aligned AGI on their own, and this is further supported by both the historical failure to align humans with arbitrary values using behavioral optimization methods and the widespread presence of Goodharting in behaviorally controlled, evolving computer systems (Scott, 1999), (Lehman et al., 2018).

Further, past research on GOFAI—AI systems based on symbol manipulation—suggests algorithmic methods of alignment are likely to be too complex to work for the same reasons that GOFAI was itself unworkable, namely that it proved infeasible for humans to program systems with enough complexity and specificity to do anything more than perform meaningless manipulations (Haugeland, 1985), (Agre, 1997). In recent years AI researchers have surpassed GOFAI

only by switching to designs where humans specify relatively simple computations to be performed and allow the AI to apply what Moravec called "raw power" to large data sets to achieve results (Russell and Norvig, 2009), (Moravec, 1976). This suggests that attempts to align AGI by algorithmic means are likely to also prove too complex for humans to solve, leaving us with only philosophical methods of alignment and thus necessitating mindful AGI.

This paints a bleak picture for the possibility of aligning mindless AGI since behavioral methods of alignment are likely to result in divergence from human values and algorithmic methods are too complex for us to succeed at implementing. This leads us to conclude that, although assuming mindful AGI has a greater risk of false positives than assuming mindless AGI all else equal, all else is not equal, mindless AGI is less likely to be successfully aligned because algorithmic and behavioral alignment mechanisms are unlikely to work, so we have no choice but to take on the risks associated with assuming mindful AGI when designing alignment schemes.

Conclusion

Based on the analysis of the previous section, we recommend AGI alignment researchers adopt moral skepticism with norm particularism and assume AGI will have minds to reduce the risk of false positives and increase the chance of success at alignment. These are not necessarily all the assumptions that must be made, but they are at least some that must be, and additional evidence and arguments may lead us to conclude that a different bundle of assumptions are better responses to the metaphysical and practical points of uncertainty that forced us to make assumptions. Future research should consider additional points of uncertainty that demand we make assumptions about AGI alignment to better understand the distribution of alignment schemes that are likely to result in acceptable outcomes for humanity.

References

Leke Adeofe. Artificial intelligence and subjective experience. In *Proceedings of Southcon '95*. IEEE, 1997.

Philip E. Agre. *Computation and Human Experience*. Cambridge University Press, 1997.

William P. Alston. Epistemic Circularity. *Philosophy and Phenomenological Research*, 47(1):1, sep 1986.

William Alston. *The Reliability of Sense Perception*. Cornell University Press, 1993.

William Alston. *Beyond Justification: Dimensions of Epistemic Evaluation*. Cornell University Press, 2005.

Stuart Armstrong. Motivated Value Selection for Artificial Agents. In *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*, 2015. Accessed on Wed, July 04, 2018.

Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. Value Alignment or Misalignment — What Will Keep Systems Accountable? In *The AAAI-17 Workshop on AI, Ethics, and Society*, 2017. Accessed on Tue, June 19, 2018.

Roy Billington and Ronald A. Allan. *Reliability Evaluation of Engineering Systems: Concepts and Techniques*. Springer, 1992.

Nick Bostrom. Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas*, Vol. 15, No. 3, 2003. Accessed on Wed, May 23, 2018.

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

Franz Brentano. *Psychology from an Empirical Standpoint*. Routledge, 1995.

Miles Brundage et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Technical report, 2018.

David Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

Roderick M. Chisholm. *The Problem of the Criterion*. Marquette University Press, 1973.

Roderick M. Chisholm. *The Foundations of Knowing*. University of Minnesota Press, 1982.

Jonathan Dancy. Ethical Particularism and Morally Relevant Properties. *Mind*, XCII(368):530–547, 1983.

Peter de Blanc. Ontological Crises in Artificial Agents' Value Systems. Technical report, 2011. Accessed on Wed, July 04, 2018.

Daniel C. Dennett. *Consciousness Explained*. Little, Brown and Co., 1991.

Hubert L. Dreyfus. *What Computers Can't Do: The Limits of Artificial Intelligence*. HarperCollins, 1978.

Charles A. E. Goodhart. Problems of Monetary Management: The UK Experience. In *Monetary Theory and Practice*, pages 91–121. Macmillan Education UK, 1984.

John Haugeland. *Artificial Intelligence: The Very Idea*. MIT Press, 1985.

Edmund Hussrl. *Ideas for a Pure Phenomenology and Phenomenological Philosophy: First Book: General Introduction to Pure Phenomenology*. Hackett Publishing Company, Inc., 2014.

Markus Lammenranta. Reliabilism Circularity, and the Pyrrhonian Problematic. *Journal of Philosophical Research*, 28:311–328, 2003.

Joel Lehman et al. The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities. Accessed on Tue, July 03, 2018, 2018.

Jan Leike and Marcus Hutter. On the Computability of Solomonoff Induction and Knowledge-Seeking. In *Lecture Notes in Computer Science*, pages 364–378. Springer International Publishing, 2015.

William MacAskill. *Normative Uncertainty*. PhD thesis, University of Oxford, 2014. Accessed on Wed, June 13, 2018.

David Manheim and Scott Garrabrant. Categorizing Variants of Goodhart's Law. Technical report, 2018. Accessed on Tue, July 03, 2018.

Hans Moravec. The Role of Raw Power in Intelligence. Accessed 2018-07-03, 1976.

Thomas Nagel. What Is It Like to Be a Bat? *The Philosophical Review*, 83(4):435, oct 1974.

J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical Physical and Engineering Sciences*, 231(694-706):289–337, jan 1933.

Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, 10(5):e1003588, may 2014.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000. Accessed on Wed, May 23, 2018.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2009.

T. M. Scanlon. 3 Rawls on Justification. In Samuel Richard Freeman, editor, *The Cambridge Companion to Rawls*, page 139. Cambridge University Press, 2003.

James C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, 1999.

John R. Searle. *Minds, Brains, and Science*. Harvard University Press, 1984.

James E. Smith and Robert L. Winkler. The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis. *Management Science*, 52(3):311–322, mar 2006.

Nate Soares and Benya Fallenstein. Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In Victor Callaghan, James Miller, Roman Yampolskiy, and Stuart Armstrong, editors, *The Technological Singularity*, page value here. Springer Berlin Heidelberg, 2017.

Nate Soares. The Value Learning Problem. In *Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence*, 2016. Accessed on Fri, June 01, 2018.

Krister Talvinen. The Inevitability of Skepticism. A Study on the Problem of the Criterion. Technical report, 2009. Accessed on Thu, May 24, 2018.

Alexey Turchin and David Denkenberger. Classification of global catastrophic risks connected with artificial intelligence. *AI & Society*, may 2018.

Amos Tversky. Intransitivity of preferences. *Psychological Review*, 76(1):31–48, 1969.

Roman V. Yampolskiy. Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach. In *Philosophy and Theory of Artificial Intelligence, SAPERE 5*. Springer-Verlag, 2012. Accessed on Wed, May 23, 2018.

Eliezer Yudkowsky. Coherent Extrapolated Volition. Technical report, 2004. Accessed on Thu, June 07, 2018.

Eliezer Yudkowsky. AI Alignment: Why It's Hard, and Where to Start. In *Symbolic Systems Distinguished Speaker Series*, 2016. Accessed on Tue, May 15, 2018.

Eliezer Yudkowsky. Goodhart's Curse. Accessed on Tue, July 03, 2018, 2017.

Eliezer Yudkowsky. Security Mindset and Ordinary Paranoia. Technical report, 2017. Accessed on Wed, May 23, 2018.

Aaron Zimmerman. *Moral Epistemology*. Routledge, 2010.