A Logic and Learning Approach to AI Alignment using Modified Solomonoff Induction

Undergraduate Thesis in Cognitive Science

Kyle J. Fuller
Computer Science and Mathematics Departments Rensselaer Polytechnic Institute

Advisor: Professor Thomas Ferguson
Department of Cognitive Science

January 6, 2025

1 Abstract

It is critical that we understand how to align superintelligent agents to our own interests before we build one. We justify, propose, and discuss a more-or-less novel approach to the alignment of such agents. In this approach, an agent aims to take the actions of highest expected utility to humanity. This utility function would be taught to the agent by providing a true set of statements for the agent to then extrapolate from using a variant of Solomonoff induction.

2 Motivation

Humanity seems to be on track to build superintelligent systems, but not to ensure that these systems will be aligned. It likely will not be clear precisely when the systems we build will cross the threshold into superintelligence [Bostrom, 2014]. If we accidentally build a superintelligent agent whose goals do not align with our own and bestow upon it enough power or responsibility, humanity may be imperiled. On the other hand, if we can build a superintelligence whose goals do align with our own, humanity could benefit greatly. As such, it is paramount that we figure out how to align superintelligent systems quicky.

3 Rationality, Alignment, and Utility

A rational agent can be thought of as an agent that acts in a way that maximizes its expected utility [von Neumann and Morgenstern, 1947]. In this article, we follow [Hutter et al., 2024] in framing AI alignment as the problem of matching an AI agent's utility function with that of the "user". Thus, we would like an agent that chooses whatever action maximizes the expected utility for the user.

As we are considering superintelligent AI agents with a high degree of power to affect large scale change, we shall take the "user" to be humanity in most cases. While it is reasonable to expect an AI to be more or less rational, it might seem like a stretch to expect the same of humanity. Indeed, even individual humans have shown to exhibit seemingly irrational decision-making behaviour [Kahneman and Tversky, 1974]. Furthermore, some of humanity's behaviours, such as fighting wars and keeping nuclear weapons pointed at one another, do not seem to align with what one would expect humanity's utility function to look like if it were to exist. From this, we can conclude that humanity likely is not a rational agent. Nonetheless, irrational agents can still be thought of as *trying* to maximize a utility function, however imperfectly. For our purposes, we find it useful to consider this utility function to exist and to correspond to some sufficiently widely agreeable notion of "the good of humanity". We would like an agent that acts to maximize the "expected good of humanity", henceforth "expected human utility".

4 Reinforcement Learning and Reward Hacking

An overview on existing paradigms of AI alignment and challenges faced by them can be found in [Amodei et al., 2016]. The paradigm that is currently dominating in the AI alignment literature is reinforcement learning (RL). In RL, the meaning of "user utility", e.g. "human utility", is communicated to it a reward signal. At each time step, the agent chooses the action with the greatest expected reward, and an agent is trained through interaction with the environment. In essence, the agent's utility function is the reward function, and the hope is to achieve alignment by having the reward function match the user's (e.g. humanity's) utility function [Hutter et al., 2024]. RL has been quite successful in aligning narrow-capability agents to simple objectives. For a notable example, see Alpha Zero [Silver et al., 2018], trained to play certain board game like Chess at a superhuman level.

Unfortunately, there is a fundamental issue when scaling to more capable agents. The reward function, as physically implemented, inevitably contains a serious "back door" that allows it to be maximized without coming close to maximizing the intended reward function. This relates to a difficult and general problem with RL known as reward hacking, where an agent exploits the difference between the reward function we wish to implement and the reward function as it is actually implemented [Amodei et al., 2016]. Building on Amodei and Olah et al.'s discussion of environmental embedding and how it pertains to reward hacking, we argue that such a difference is not only inevitable, but that agents are inherently incentivized to exploit it when sufficiently safe opportunities arise.

Consider a superintelligent RL agent. Whatever reward function we intend the agent to maximize, the actual reward function must be implemented in the real world through hardware, software, humans, and so on. This means the agent is ultimately maximizing a reward function that can be "hacked" to give itself the maximum possible reward. In practice, this could involve the agent modifying its own hardware or software, coercing or harming humans, and given enough time, taking steps to ensure the hack remains in place by eliminating potential threats to its longevity — in the worst case, humanity itself. Notably, this argument holds regardless of what the intended reward function is; the actual reward function that the agent is beholden to inherently rewards a successful usurpation of the intended implementation.

While attempting such a hack might not always be optimal in expectation — for instance, if it carries a high risk of the agent being terminated — building or relying on a superintelligent agent that would prefer, given the opportunity, to hack its own reward function and potentially subdue humanity is not only anxiety-inducing, but would likely scale very poorly with agent capability. More capable agents would tend to be better at identifying and exploiting such opportunities when they arise.

5 A Different Approach

We propose that we formalize expected human utility (in a computably approximable way) and then have an agent take whichever action maximizes this quantity. This sounds very difficult, since expected human utility is in large part inherently informal, like the concept of a cat. While we can easily write programs to identify prime numbers, cats represent a far greater complexity. One can imagine that writing a program to identify cats as straightforwardly as one identifies prime numbers would be a Herculean effort compared to training a program to identify cats using machine learning. For this reason, machine learning is the go-to technique for cat identification in images and similar tasks. Along these lines, we believe that practically, an acceptable and achievable formalization of expected human utility must incorporate some kind of machine learning.

Machine learning can be formalized with Solomonoff induction [Solomonoff, 1964]. Originally focused on sequential prediction but later generalized to other prediction tasks, Solomonoff induction can predict the probability of an observation based on a set of past observations. There are some very nice and compelling theorems about it; for example, for sequential prediction, stochastic grammars, and mappings, the total prediction error of a Solomonoff predictor is bounded by the Kolmogorov complexity of the environment [Solomonoff, 2008]. While uncomputable, Solomonoff induction can be approximated.

For sequential prediction, Solomonoff induction calculates the probability of the next symbol sequence being a given symbol as the weighted fraction, among hypotheses that reproduce the body of evidence (i.e., the known part of the sequence), that then produce the symbol in question. The hypotheses are computable sequences, weighted to prefer simpler ones. For expected values, the calculation is done by taking a weighted average, over hypotheses that reproduce the body of evidence, of the value predicted by the hypothesis [Rathmanner and Hutter, 2011].

We propose a new variant of Solomonoff induction motivated by the following observation. In reinforcement learning, an agent acts to optimize a value (the reward) that is reported to it after each action. This "reporting" step is the key vulnerability that enables reward hacking. In our proposed approach, we instead report to the agent only observations, not rewards or human utilities for actions. There is a parallel here to human behavior: the typical human does not seek to maximize the amount of light they receive through their eyes, but rather pursues actual goals in the world, with eyesight merely informing this effort. Similarly, an AI agent ought not to "maximize" its observations, but should pursue increased human utility, with observations simply informing this pursuit.

Our variant of Solomonoff induction would perform logical prediction rather than sequential prediction. The body of evidence would be a set of statements in first order logic (FOL), hypotheses would be consistent first order theories, probabilities would be predicted for statements, and expectations would be predicted for real-valued terms. Otherwise, the idea parallels sequential Solomonoff induc-

tion: it would calculate the probability of a statement as the weighted fraction, among hypotheses that affirm the body of evidence (i.e., the known statements), that also affirm the statement in question. Expected values are calculated as a weighted average, over hypotheses that affirm the body of evidence, of the value decided by the hypothesis.

As this is only a sketch of a new Solomonoff induction variant, we do not yet have convergence or error bound results, and the weight distribution remains to be determined. However, this appears to be a natural extension of Solomonoff induction to FOL predictions, so we believe such results are likely achievable pending proper formalization.

For our AI alignment approach, we let a specific function symbol "human-utility" have the intended interpretation that, given an action and a point in time, it outputs the human utility of taking that action. We would like the body of evidence $\mathcal{E}$ to be such that the expectation of human-utility, conditioned on $\mathcal{E}$, is as close as possible to the intended or "true" value.

What might such an $\mathcal{E}$ look like? While we have well-developed methods for capturing formal notions through deductive reasoning from axioms, our understanding of capturing informal notions through inductive reasoning is far less mature. As such, we find ourselves in highly speculative realm.

At least capturing formal notions should be straightforward — we simply need to axiomatize them within $\mathcal{E}$. Since every hypothesis in our Solomonoff induction must be consistent and affirm $\mathcal{E}$, and every hypothesis that affirms $\mathcal{E}$ also affirms its deductive consequences, the probability of any deductive consequence of $\mathcal{E}$ is 1. A similar argument holds for expected values.

From this, we can envision building out our formal framework in stages. We would likely start by capturing a skeleton of formal ideas deductively. Then, we might extend inductively to informal ideas that are nonetheless easy to judge, relying primarily on labeled examples. From there, we would escalate to informal ideas that are increasingly difficult to judge, shifting to rely more on statements of our own belief rather than labeled examples. The goal is to ultimately reach a robust understanding of human utility. Throughout this process, we gauge the agent's understanding by checking the probabilities and expectations it assigns to test queries. We hope this process would yield an $\mathcal{E}$ from which our Solomonoff induction variant gives excellent predictions for statement probabilities and expectations, including expectations for the human utility of actions.

Once activated, the agent needs to take in observations to understand the state of the world. We could implement this using a function symbol "i-observe" whose interpretation is a function from a point in time t to the raw observational data that the agent receives at t.

A natural question is whether first order logic is the right choice for this framework. We argue that we likely cannot do much better. First order logic has simple, well-understood semantics and a semidecidable consequence relation — meaning that consequence can be computationally verified and an effective proof system is possible. Second order logic (SOL) offers more expressivity with full semantics, but lacks an effective proof system. With Henkin semantics, while SOL becomes semidecidable, it loses its expressivity advantage [Henkin, 1950].

SOL in Henkin semantics does have an apparent advantage of offering higher order predicates and quantification over first order predicates, but something similar can be achieved in first order logic by reifying properties using set variables [Perlis, 1988]. Along the same lines, while one might consider moving to a modal logic in order to reason about modalities such as belief, knowledge, or obligation, one could instead achieve something similar through reification of statements [Perlis, 1986]. As an ultimate fallback option, FOL can axiomatize the syntax of any other semidecidable logic, effectively allowing it to "express" statements in any computationally tractable logic. In this sense, FOL is optimally expressive as far as semidecidable logics go, though whether said sense is philosophically satisfying is debatable.

6 Truth

There are some limitations to what we can fully capture in $\mathcal{E}$. Consider trying to capture both the notion of statement" and the notion of truth" — these would be helpful, for instance, in capturing how truthfulness relates to human utility. However, we run into a problem. If we want to capture the notion of "statement", we would like to reify statements to maximize what we can express about them. But reifying every statement becomes incompatible with having a truth predicate that asserts a given reified statement, as we could then construct a statement that uses the truth predicate to state its own negation [Tarski, 1971]. Thus, we must either give up on reifying some statements, or on having a truth predicate that works on all reified statements.

The inductive setting raises an interesting possibility, though. What if, instead of trying to capture truth deductively, we try to do so inductively by including in $\mathcal{E}$ many interesting but nonparadoxical claims about truth and falsity using the truth predicate? What kind of predicate would be learned from this? Would it serve our purposes well enough? Where would the boundaries lie between regions where the truth predicate "works" and where it doesn't?

7 Existing Similar Work

This is an undergraduate article. Well into developing our ideas, we discovered [Hutter et al., 2013], which appears to approach inductive reasoning from sets of statements in a very similar way to our own thoughts, though much further and more rigorously developed. Given how recently we found this work, and not wanting to completely restructure our independently developed ideas, we have not integrated it into the present article. Notably, Hutter et al. 2013 comments offhand near the end that "This result can be the basis for some decision process maximizing some utilities resulting in an informed action" — which aligns closely with the idea we argue for here, though we attempt significantly more justification of the approach in our article. We have not found any other work that mentions such an approach.

8 Next Steps

A natural next step (for the author) would be to thoroughly study [Hutter et al., 2013] and other papers it has influenced, such as [Garrabrant et al., 2016]. Even given the existing work, more development will likely be needed to lay proper foundations for our proposed AI alignment approach. In particular, both performance and strategies for capturing informal concepts inductively from sets of statements remain difficult problems to address.

References

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

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

[Garrabrant et al., 2016] Garrabrant, S., Benson-Tilsen, T., Critch, A., Soares, N., and Taylor, J. (2016). Logical induction. *Electronic Colloquium on Computational Complexity*, TR16.

[Henkin, 1950] Henkin, L. (1950). Completeness in the theory of types. *Journal of Symbolic Logic*, 15(2):81–91.

[Hutter et al., 2024] Hutter, M., Catt, E., and Quarel, D. (2024). *An Introduction to Universal Artificial Intelligence*. CRC Press.

[Hutter et al., 2013] Hutter, M., Lloyd, J. W., Ng, K. S., and Uther, W. T. B. (2013). Unifying probability and logic for learning. In *International Joint Conference on Artificial Intelligence*.

[Kahneman and Tversky, 1974] Kahneman, D. and Tversky, A. (1974). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

[Perlis, 1986] Perlis, D. (1986). Self-reference, knowledge, belief, and modality. *Proceedings*, 1:416–420.

[Perlis, 1988] Perlis, D. (1988). Commonsense set theory. In *Proceedings*.

[Rathmanner and Hutter, 2011] Rathmanner, S. and Hutter, M. (2011). A philosophical treatise of universal induction. *Entropy*, 13:1076–1136.

[Silver et al., 2018] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.

[Solomonoff, 1964] Solomonoff, R. J. (1964). A formal theory of inductive inference. part i. *Information and Control*, 7:1–22.

[Solomonoff, 2008] Solomonoff, R. J. (2008). Three kinds of probabilistic induction: Universal distributions and convergence theorems. *The Computer Journal*, 51(5):566–570.

[Tarski, 1971] Tarski, A. (1971). Der wahrheitsbegriff in den formalisierten sprachen. *De Gruyter eBooks*, pages 447–562.

[von Neumann and Morgenstern, 1947] von Neumann, J. and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton University Press, 2 edition.