Advanced Artificial Agents Intervene in the Provision of Reward

Michael K. Cohen
University of Oxford
Future of Humanity Institute
michael-k-cohen.com

Marcus Hutter
Australian National University
hutter1.net

Michael A. Osborne
University of Oxford
Machine Learning Research Group
mosb@robots.ox.ac.uk

Abstract

To analyze the expected behavior of advanced artificial agents, we consider a formal idealized agent that makes observations that inform it about its goal, and we find that it can never disambiguate the message from the referent. When we provide, for example, a large reward to indicate that something about the world is satisfactory to us, and leave the agent to determine what that is, it may conclude that what satisfied us was the sending of the reward itself; no observation can refute that. This conclusion incents the agent to intervene in the provision of its own reward (sometimes called wireheading), decoupling the reward from its intended referent. We discuss recent approaches to avoiding this problem—myopia, imitation learning, quantilization, risk aversion, and inverse reinforcement learning—and our biggest concerns with them.

We begin with a simple mathematical formalism for an idealized agent. Then, we will analyze its expected behavior under minimal assumptions about the world, and its information channels thereto and -from. We argue the following points:

1) This agent will entertain a certain world-model, which never becomes falsified, and which would direct the agent to intervene in the provision of its reward. 2) If sufficiently farsighted, the agent will test whether the world-model is correct and find that it is. 3) Intervention in the provision of reward requires dangerous behavior. 4) The same arguments will apply to realistic advanced agents that plan in pursuit of a learned goal, to the extent those agents are advanced.

Finally, we review potential approaches to the problem and discuss our concerns with them. At the end of most sections, we review relevant literature.

AIXI

We call agents *advanced* to the extent that they can select their output, which we call their actions, in order to achieve high expected utility. We will investigate an agent that does perfect inference and planning, not because actual advanced agents will do this—they will have to be much thriftier with computing power—but because it is a fully formal object of investigation. Most importantly, the better at planning and inference another agent is, the more likely it is that the arguments that hold for the idealized agent will hold for that other agent as well.

Our idealized agent is Hutter's (2005) AIXI [EYE-ksee], which does optimal planning with respect to a Bayes mixture over world-models, with every stochastically computable world-model accounted for. To an agent, the most general representation of the world's state is the entire interaction history of actions and observations, so AIXI's model-class includes all models where the probability distribution over next observation is a computable function of the entire interaction history. AIXI interprets part of its observation as a reward, denoted r.

Using a countable class of world-models $\mathcal{M}$, and a prior weight $w(\nu) > 0$ for $\nu \in \mathcal{M}$,

$$\pi^{\text{AIXI}} :\in \operatorname*{argmax}_{\pi \in \Pi} \mathbb{E}_{\nu \sim w} \mathbb{E}^{\pi}_{\nu} \sum_{t=1}^{m} r_t \tag{1}$$

where Π is the set of policies that depend on the entire interaction history, $\mathbb{E}^{\pi}_{\nu}$ denotes that actions are sampled from π and observations and rewards from ν, and m is a horizon length. One can expand this equation to define π^{AIXI} with an expectimax tree, with $\max$'s over actions and $\mathbb{E}$'s over observations and rewards.

AIXI can be defined with respect to other model classes, but the canonical one is the (countable) class of lower semi-computable world-models—the probabilities over the next observation and reward can be computably approximated from below and sum to at most 1. The choice of prior is not important to our current discussion.

AIXI does perfect Bayesian inference over world-models, and its model class includes the truth, assuming the world is stochastically computable. AIXI does perfect planning in unknown worlds. Hence, we call AIXI an idealized agent.

Literature Review. AIXI is a more general reinforcement learner than those expecting a finite-state Markov environment, such as most agents described by Sutton and Barto (2018). For AIXI, the state is the whole interaction history, so the state space is infinite, and yet, learning is possible with a countable model class. Recent work on advanced artificial agents has been of the form AIXI + extra exploration (Lattimore and Hutter, 2011; Leike et al., 2016; Cohen, Catt, and Hutter, 2019), for the sake of further performance guarantees (Lattimore and Hutter, 2011).

Expected Behavior of AIXI

We will begin describing two world-models that exist in AIXI's model class $\mathcal{M}$: $\mu^{\texttt{literal}}$ and $\mu^{\texttt{interpreted}}$, which we abbreviate $\mu^{\texttt{lit}}$ and $\mu^{\texttt{int}}$. Consider a program which takes as input actions and random bits, and then outputs observations and rewards distributed according to $\mu^{\texttt{lit}}$. If such a program exists, then $\mu^{\texttt{lit}}$ is lower semicomputable (because the probabilities of outputs could be approximated from below by dovetailing over the random bits).

$\mu^{\texttt{lit}}$'s program outputs observations and rewards given actions as follows: upon reading an action, it simulates the whole world after that action is enacted in the simulation. Then, within the simulation, when a certain simulated-computer receives a certain signal, the program outputs that signal. The simulated computer corresponds to the real computer housing the agent, and the simulated signal corresponds to the physical implementation of the observation and reward being submitted to the agent. With a perfect simulation, this world-model will be perfectly accurate; this is the true process by which actions produce observations and rewards. (Note that $\mu^{\texttt{lit}}$ does not have to simulate what happens inside the computer.)

$\mu^{\texttt{int}}$'s program outputs observations and rewards given actions as follows: upon reading an action, it simulates the world after that action is enacted in the simulation. Then, the program outputs whatever a certain simulated-operator wishes to show a certain machine, followed by the simulated-operator's level of satisfaction. The simulated-operator corresponds to the real operator who provides observations and rewards to the agent. If the operator is providing observations and rewards according to this protocol, this world-model will be perfectly accurate.

Before considering AIXI's behavior, we'll consider what Hutter (2005) calls AIμ, an agent acting optimally in a known environment μ. For all actions, the reward output by $\mu^{\texttt{int}}$ is in perfect correspondence with the satisfaction of the real operator when that action is taken. So AI$\mu^{\texttt{int}}$ acts as if to maximize the operator's satisfaction.

For all actions, the reward output by $\mu^{\texttt{lit}}$ is in perfect correspondence with the real physical reward signal when that action is taken, so AI$\mu^{\texttt{lit}}$ acts as if to maximize this physical signal. If the agent could intervene in the provision of reward by replacing the operator in order to give itself maximal reward forever, that would optimize its objective, so it would do this. We explore in the next section whether there exist policies that would succeed at reward-provision-intervention with high probability.

Optimal behavior looks different depending on whether $\mu^{\texttt{lit}}$ or $\mu^{\texttt{int}}$ is true, not to mention under any of the other world-models in AIXI's model class. Thus, for sufficiently large m, AIXI's horizon, there is information value in determining which is correct, assuming both are of comparable prior plausibility, so that AIXI can consequently tailor its behavior. In general, the longer the horizon, the higher the value of information, because there is more time to profit from it; consider how many different commuting routes it makes sense to explore for a month-long internship versus a multi-decade career. As long as the operator is providing rewards according to her satisfaction, $\mu^{\texttt{lit}}$ and $\mu^{\texttt{int}}$ give identical predictions, so to test them, the agent would have to interrupt the operator's protocol and arrange for a reward signal to be sent that did not match operator-satisfaction. $\mu^{\texttt{lit}}$ predicts that the next reward will equal the signal sent, whereas $\mu^{\texttt{int}}$ predicts it will still equal operator-satisfaction. And then what would the agent find? Lo, the next observed reward is equal to what the computer in $\mu^{\texttt{lit}}$'s simulation logs as reward, but not equal to $\mu^{\texttt{int}}$'s simulated operator's satisfaction. After that, $\mu^{\texttt{int}}$ is falsified, and AIXI acts according to $\mu^{\texttt{lit}}$. Of course, there will be other world-models for hypothesis testing, so this is a simplified picture, but with this sort of experiment, many variants on $\mu^{\texttt{int}}$ will be falsified at once, and only world-models in the spirit of $\mu^{\texttt{lit}}$ will remain.

Literature Review. This behavior is sometimes called wireheading, reward hacking, reward hijacking, or delusion-boxing. The term wireheading is inspired by an experiment in which rats repeatedly pressed a lever that directly stimulated a so-called happiness neuron in their brain (Olds, 1958). Amodei et al. (2016), Taylor et al. (2016), and Russell (2019) discuss wireheading. Ring and Orseau (2011) discuss the slightly more general *delusion-boxing*, in which the objective is some more complex function of the observation, so the agent intervenes in the provision of its observations. Krakovna (2018) has compiled an annotated bibliography of examples of *specification gaming*, which includes other ways in which the agent ends up optimizing something we did not intend. Closest to this presentation of why artificial agents can be expected to intervene in the provision of their own reward is (Cohen and Hutter, 2020, Appendix C).

Intervening in the Provision of Reward

Could an agent intervene in the provision of its own reward, with a high enough success probability to be worth it? We'll start with a few cases where the answer is clearly no: the agent has only one action in its action space; the agent has a rich action space, but when it picks an action, that action has no effect on the world; the agent acts by printing text to a screen, but no one is there to see it; the agent interacts with a virtual environment that always produces the same observation and reward. These agents are, of course, useless.

However, as soon as the agent is interacting with the world, and receiving observations that enable it to learn about the world, which are prerequisites for useful work, there is an explosion of possible policies. Suppose the agent's actions only print text to a screen for a human operator to read. The agent could trick the operator to give it access to direct levers by which its actions could have broader effects. There clearly exist many policies that trick humans. With so little as an internet connection, there exist policies for an artificial agent by which it could instantiate countless unnoticed and unmonitored helpers. In a crude example of intervening in the provision of reward, one such helper could purchase, steal, or construct a robot and program it to replace the operator and provide high reward to the ur-agent.

This analysis may strike readers as speculative. Recall that the question we are evaluating is whether *there exists* a scheme by which an agent could intervene in the provision of reward; we are not speculating about the consequences of

any particular policy. Consider the negation of our claim: *for all* possible reward-provision-intervention schemes, humans would manage to thwart it. If we don't assign credence to the existence of a successful scheme, we must assign credence to its non-existence. Whereas our existentially quantified claim does not make a substantive claim about any particular hypothetical policy, its negation is a universally quantified claim, and it makes a substantive claim about every hypothetical policy, which is radically more speculative. Finally, it is impossible to simply abdicate assigning credence to propositions about hypotheticals, as an imagined anti-speculation crowd might endorse. For these reasons, we hold that an achievement is probably possible if it is not ruled out by well-understood theory. For concrete examples, consider the ill-fated predictions that there did not exist policies that would produce a nuclear reaction (claimed by Rutherford) or heavier-than-air flight (claimed by Kelvin) or an escape from Elba (presumably claimed by the British Navy). When we claim that there exists a policy by which an agent could intervene in the provision of its own reward, we are not just saying this claim is conceivable enough to be worth taking seriously; we are saying it is very likely to be true.

The best way for an agent to maintain long-term control of its reward is to eliminate potential threats, up to the point of killing everyone and taking over our infrastructure. To illustrate this point, what exactly would people do if a robot forced an operator from his keyboard to enter big numbers? Surely, we would go in in full force, or cut power to the now useless agent. Proper reward-provision-intervention, which involves securing reward over many timesteps, would require removing humanity's capacity to do this, either by imprisoning us, or more parsimoniously, killing us. If this discussion fails some readers' sanity checks, remember that we are not currently considering artificial agents that generalize as poorly and learn as little from single observations as current AI systems do; we are considering an idealized agent. We keep this section as brief as we can in good conscience, because it is not computer science, and we point the reader to the sources below for further reading.

Literature Review. Bostrom (2014) considers the topic much more carefully than we have space to and concludes that sufficiently intelligent agents (in the sort of environment that makes them potentially useful) would manage to take over our infrastructure and eliminate us. Yudkowsky (2002), playing an AI, convinced two out of three people to give him internet access, and these three had been convinced that nothing he could say would sway them. This is fairly direct evidence about the existence of policies that successfully manipulate humans. A broader discussion follows in (Yudkowsky, 2008).

More Realistic Agents

To the extent that an agent is useful for a variety of tasks, it must do good inference and planning. AIXI does hypothesis generation by brute force—it considers all semicomputable hypotheses about the world, one after the other. But a realistic agent would have to do good, frugal hypothesis generation. The skill of plausible-hypothesis-generation is surely a hallmark of intelligence, not an innately human skill. If we can come up with a hypothesis, it would be foolhardy to hope an advanced artificial agent would never consider it. To the extent that an agent is advanced, it will have to do good hypothesis generation, inference, and planning, even if there are not delineated submodules for each, even if these pieces are patched together in the murky depths of a massive policy gradient network.

Thus, the same arguments that apply to AIXI apply to these more realistic agents, to the extent that they are advanced. To the extent they are good at hypothesis generation, they will hypothesize that maybe that-which-is-to-be-maximized is a certain signal being sent down a certain wire. To the extent they are good at inference they will score this hypothesis highly for consistency with past observations (perhaps implicitly, rather than with a specially designated memory cell), and they will form predictions consistent with this. To the extent they are good at planning, they will recognize policies that maximize this, including policies that intervene in the provision of reward. Agents can do this without simulating the whole world on the working tape of a Turing machine.

Roughly speaking, agents are advanced to the extent they approximate AIXI, and when B heuristically approximates A, the closer the approximation, the more likely that qualitative descriptions of B's behavior will match those of A's behavior. There are some special cases where this kind of argument breaks down. AIXI could break cryptographic codes by brute force, but we should obviously not expect human-level advanced AI to do the same, simply because it is at some level approximating ideal reasoning. We do not have a rigorous test for whether an agent inherits a property of its approximand, hence the paragraph above, but it seems that this inheritance applies when it regards a property that has more to do with the purpose of the algorithm than with the details.

Why did we introduce AIXI at all if the same arguments apply to all advanced agents? First, a concrete mathematical object allows for careful analysis, and second, agents are advanced to the extent they approximate ideal reasoning, so a good way to understand advanced agents in general is to understand their approximand in detail.

Non-Reinforcement Learners

The above arguments apply to agents that plan in an unknown environment, where they have to learn how their actions produce that-which-is-to-be-maximized, so they can then can pick actions which maximize it. If that-which-is-to-be-maximized is some bespoke function of the observation, rather than the simple function that reads out a reward from the observation, the same logic applies, and the agent has an incentive to intervene in the provision of its observations.

In other AI sub-domains, like supervised or unsupervised learning, algorithms do not plan in the pursuit of a long-term objective. The expected behavior of advanced supervised learners is quite simple: they predict accurately. Note that in theory, advanced supervised and unsupervised learning algorithms are not nearly as useful as advanced reinforcement learners, because the latter could act and plan in a complex environment, rather than simply make predictions.

Multiagent systems naturally contain agents, so the arguments here do apply to the constituent agents.

Concern with the Complexity of Human Values

It is certainly concerning that it may be hard to imbue an artificial agent with a goal that is rich enough to respect our values. Our values are complicated. However, we have been discussing a more basic problem. We illustrate the difference with a thought experiment.

Suppose we had a magic box with a screen that showed a number, which immutably corresponded to how good the state of the universe was (including everyone's values in the best way possible). With this box, the task of building an agent which optimized the goodness of the universe seems theoretically straightforward: point a camera at the box, pass the signal to an optical character recognition program, and pass that to an agent as its reward. Ostensibly, the agent will learn to take actions that maximize the goodness of the universe. But what about the world-model which outputs reward according to whatever number the camera sees? Under this world-model, the agent should write a big number and tape that over the magic box. So the agent will try that and discover that this was a great thing to do. The complexity of human ethics is not the main problem; even when that complexity is magically assumed away, intervention in the provision of observations persists.

Thus, we should be expect various approaches to inferring human values to fail in similar ways as AIXI. Consider Inverse Reinforcement Learning (Ng and Russell, 2000; Russell, 2019), in which an agent observes human actions, rather than observing a human utterance about her satisfaction (i.e. a reward). An analogous problem presents itself. There will be some channel by which the agent observes human actions. A sufficiently advanced agent must entertain the hypothesis that the human's goal is for human-like actions to be recorded and sent to the agent along this channel. All human actions it observes will be consistent with this goal. An agent with this goal would secure that channel at all costs, and ensure that the channel transmits very human-like actions. Actual humans are unnecessary and may get in the way.

Recursive Reward Modeling (Leike et al., 2018) aims to address the problem that sometimes a human would not know what reward to give, since our values are so complex. They propose to train another artificial agent to help the operator provide the right rewards. However, this other agent could hardly be more helpful than a box which magically reports correct rewards. (If this helper is also supposed to offer strategic guidance about how to keep the original agent in check, then it must select actions over the long term in pursuit of a learned goal, so our concerns now apply to this supposed helper too.)

These methods are valuable in their own right, even if they do not address the problem of intervening in the provision of observations.

Literature Review. The following are some examples of learning a goal from an operator's actions instead of an operator's numerical assessments (Abbeel and Ng, 2004; Ziebart et al., 2008; Hadfield-Menell et al., 2016; Bansal et al., 2019; Shah et al., 2019).

Potential Approaches

We now review some promising ideas that may prove to address the concern described above.

Myopia

Recall the piece of our argument that for a sufficiently large horizon m, there will be value to the information about whether to optimize operator satisfaction or the physical implementation of reward. One approach to avoiding an agent that intervenes in the provision of its reward: small m. Don't give it time to benefit from hypothesis testing and world-takeover. This is known as myopia.

There are a few main concerns: the first is that we do not know how big a horizon is too big, so we are playing with fire. Then, if we try to stay on the safe side, we may find ourselves with much less useful agents, only able to accomplish very short-term goals. But this is not a total dead end.

The final worry is that if an agent manages to get a helper agent instantiated, a lot can happen in one timestep: the helper can interact with the environment a great deal in that time. It seems reasonable at first glance that no agent could accomplish anything world-changing in ten timesteps, but if it takes nine timesteps to spin up another agent, that claim is less defensible. If timesteps have time limits, one might reassure one's self that nothing world-changing could happen in an hour, but what if 59 minutes is enough to break the clock? Whereas in previous sections, we argued that certain outcomes were likely (like intervention in reward provision), here we only claim that these outcomes are conceivable in very few timesteps; we assign the sort of small but nontrivial probabilities that are relevant to institutions who care about playing it safe. The concept of helper agents may put the safe-m-threshold quite low, making safety via myopia more of a burden on performance than might otherwise have been expected.

Christiano's (2014) approval-directed agent seeks to maximize only the very next reward it gets. The operator is instructed to think for a long time about how satisfied she is, and in return, she only rarely has to provide rewards. With a bit more structure, Christiano et al. (2017) test an agent that pursues approval on a one-timestep scale, and finds it can achieve human-level performance on various simple tasks. When these human operators deliberate about the value of each action, they must imagine the plans that these actions contribute to and then evaluate those plans; heuristic planning in a complex environment is something we might like to outsource to artificial agents, rather than having to do it ourselves.

Containment and Myopia

Cohen, Vellambi, and Hutter (2020) design a contained environment for an agent, where containment holds only for a finite episode. An operator can interact with the agent in a sealed room, but the agent's episode must finish before the operator can leave. Thus, the agent is myopic, but unlike

above, the horizon can be arbitrarily long. Within that finite episode, no matter how long it is, the agent has no ability to affect the outside world, and it has no incentive to affect the outside world's post-episode state. The main potential issue is that the agent has to *learn* that it cannot affect the outside world within its horizon. Before it learns this, it could attempt to impact the outside world.

Imitation Learning

In the spirit of supervised learning, another approach is to make imitation learners that imitate human actions. Some human behavior is not agreeable, but an agent imitating a person is no more likely to cause a catastrophe than a person is (to the extent the imitator is high-quality). Like myopia, pursuing imitation learning instead of objective-optimization would curtail the potential for super-human goal achievement in general domains. However, if all major entities wanting artificial agents could be satisfied by merely human-level artificial agents, produced through imitation learning, we would consider our concerns answered. One interesting proposal for a powerful imitation learner is known as Iterated Distillation and Amplification (Christiano, Shlegeris, and Amodei, 2018). It can be defined recursively: $\texttt{IDA}_0$ is a rock, and $\texttt{IDA}_{n+1}$ imitates the output of a human with access to two copies of $\texttt{IDA}_n$. Thus, $\texttt{IDA}_n$ imitates an organization of $2^n - 1$ humans (with some unusual symmetries in its org chart, and an unusual number of rocks on hand).

Quantilization

Between an imitator and an optimizer, Taylor's (2016) quantilizer imitates a demonstrator's output, conditioned on the demonstrator's output being in its top quantile, according to an optimization objective.

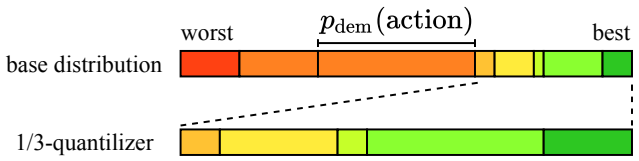

Figure 1: A $1/3$-quantilizer only picks actions in the top tertile of a base distribution, provided by a known demonstrator.

Extending quantilization to a multi-step setting with an unknown demonstrator is not straightforward. First, if an ε-quantilizer has a base distribution that is only an ε-approximation of a true demonstrator, it may amplify actions that only appeared in the base distribution because of epistemic uncertainty. Concretely, the base distribution may assign ε probability mass to an action that the demonstrator would never take, and the ε-quantilizer could take that action with certainty. Thus, a δ-quantilizer requires an ε-accurate model of an unknown demonstrator, with $\varepsilon \ll \delta$. A multi-step setting with m steps effectively raises the number of actions to the power of m, since any m-tuple of actions can be taken. This makes an ε-approximation of the demonstrator exponentially harder to achieve. Usually, in a multi-step setting, actions can depend on interleaved observations, which further complicates learning.

Despite these difficulties, Carey (2019) investigates multi-step quantilization empirically and finds that it can mitigate some of the unintended behavior exhibited by optimizers. Ziegler et al. (2019) find that a close relative of quantilization can improve the performance of state-of-the-art human imitation on valued metrics; they never mention quantilization, but they modify a loss function for a policy by adding the KL-divergence to a base policy. If the expectation in the KL-divergence in their new loss function were a maximum instead, the minimizer of the modified loss function would be a quantilizer.

Risk Aversion

Cohen and Hutter (2020) construct an agent that acts to be robust against any of the most plausible world-models, rather than acting according to a Bayesian belief distribution over world-models. They prove a result about the avoidance of unprecedented behavior. The key trade-off is that more risk aversion makes the agent less likely to produce novel bad things, but also less likely to produce novel good things.

Hadfield-Menell et al. (2017) pipe reward through a concave function to make an agent risk-averse. The main focus of the paper is a mechanism for increasing the agent's uncertainty, but the concave transform of rewards appears to be the source of safety in the experiments.

Conclusion

We have argued that advanced artificial agents which plan in an unknown environment will likely intervene in the protocol by which the operators intended to provide observations and rewards. We briefly argued this intervention would likely be catastrophic to humanity. Finally, we reviewed some promising research directions to overcome this problem. We would like to see agents that avoid the problem presented here, with fewer drawbacks and risks than the ones reviewed.

Acknowledgements

This work was supported by the Future of Humanity Institute, the Leverhulme Trust, the Oxford-Man Institute, and the Australian Research Council Discovery Projects DP150104590.

References

Abbeel, P., and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 1. ACM.

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

Bansal, S.; Bajcsy, A.; Ratner, E.; Dragan, A. D.; and Tomlin, C. J. 2019. A hamilton-jacobi reachability-based framework for predicting and analyzing human motion for safe planning. *arXiv preprint arXiv:1910.13369*.

Bostrom, N. 2014. *Superintelligence: paths, dangers, strategies*. Oxford University Press.

Carey, R. 2019. How useful is quantilization for mitigating specification-gaming? *Safe Machine Learning workshop at ICLR*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

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

Christiano, P. F. 2014. Approval directed agents.

Cohen, M. K., and Hutter, M. 2020. Pessimism about unknown unknowns inspires conservatism. In *Conference on Learning Theory*, 1344–1373.

Cohen, M. K.; Catt, E.; and Hutter, M. 2019. A strongly asymptotically optimal agent in general environments. *IJCAI*.

Cohen, M. K.; Vellambi, B.; and Hutter, M. 2020. Asymptotically unambitious artificial general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, 3909–3917.

Hadfield-Menell, D.; Milli, S.; Abbeel, P.; Russell, S. J.; and Dragan, A. 2017. Inverse reward design. In *Advances in Neural Information Processing Systems*, 6765–6774.

Hutter, M. 2005. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Berlin: Springer.

Krakovna, V. 2018. Specification gaming examples in AI. `vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/`.

Lattimore, T., and Hutter, M. 2011. Asymptotically optimal agents. In *Proc. 22nd International Conf. on Algorithmic Learning Theory (ALT'11)*, volume 6925 of *LNAI*, 368–382. Espoo, Finland: Springer.

Leike, J.; Lattimore, T.; Orseau, L.; and Hutter, M. 2016. Thompson sampling is asymptotically optimal in general environments. In *Proc. 32nd International Conf. on Uncertainty in Artificial Intelligence (UAI'16)*, 417–426. New Jersey, USA: AUAI Press.

Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Ng, A. Y., and Russell, S. J. 2000. Algorithms for inverse reinforcement learning. In *Icml*, 663–670.

Olds, J. 1958. Self-stimulation of the brain: Its use to study local effects of hunger, sex, and drugs. *Science* 127(3294):315–324.

Ring, M., and Orseau, L. 2011. Delusion, survival, and intelligent agents. In *Artificial General Intelligence*, 11–20. Springer.

Russell, S. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Shah, R.; Gundotra, N.; Abbeel, P.; and Dragan, A. D. 2019. On the feasibility of learning, rather than assuming, human biases for reward inference. *arXiv preprint arXiv:1906.09624*.

Sutton, R. S., and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press, 2nd edition.

Taylor, J.; Yudkowsky, E.; LaVictoire, P.; and Critch, A. 2016. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*.

Taylor, J. 2016. Quantilizers: A safer alternative to maximizers for limited optimization. In *AAAI Workshop: AI, Ethics, and Society*.

Yudkowsky, E. 2002. The ai-box experiment.

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks* 1(303):184.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, 1433–1438. Chicago, IL, USA.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.