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
Google DeepMind
Australian National University
hutter1.net

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

Abstract

We analyze the expected behavior of an advanced artificial agent with a learned goal planning in an unknown environment. Given a few assumptions, we argue that it will encounter a fundamental ambiguity in the data about its goal. For example, if we provide a large reward to indicate that something about the world is satisfactory to us, it may hypothesize that what satisfied us was the sending of the reward itself; no observation can refute that. Then we argue this ambiguity will lead it to intervene in whatever protocol we set up to provide data for the agent about its goal. We discuss an analogous failure mode of approximate solutions to assistance games. Finally, we briefly review some recent approaches that may avoid this problem.

We call an agent *advanced* to the extent that it effectively selects its output, which we call its actions, in order to achieve high expected utility in a broad set of environments. Since we will likely want advanced artificial agents to operate in environments for which we lack the source code, like the real world, we consider agents acting in an environment that is unknown to them. If the agent's goal is not simply a hard-coded function of its actions, then it must depend on the agent's percepts too. Percepts that indicate goal-attainment essentially inform the agent that somehow, whatever it has made happen is good. Thus, our inquiry regards agents that plan actions in an unknown environment, which requires them to learn which actions serve their goal.

We begin with an idealized situation, in which we appear to have all the tools we need to create an advanced agent with a good goal. We identify a key ambiguity the agent faces, which we argue will likely motivate the agent to intervene in the protocol by which we intended to provide goal-informative percepts. We then generalize the argument to other situations with reward-based goal-information. Rewards are percepts that quantify how well the goal has been achieved. (Often, an agent that gets reward also gets another percept called an observation that simply helps it learn a model of the world). As a sanity check, we confirm that these arguments apply to an idealized artificial agent that does perfect reasoning under uncertainty and perfect planning, this being the limit of advancement. Next, we argue that an advanced agent motivated to intervene in the provision of reward would likely succeed and with catastrophic consequences. We then discuss how a similar failure mode faces an artificial agent in an assistance game (Hadfield-Menell et al., 2016). Finally, we discuss potential approaches that may undermine the assumptions of our argument.

Related Work

We are not the first to expect reinforcement learners to intervene in the provision of reward, but we are unaware of other work that explicitly lays out a set of assumptions from which that follows. And we ultimately generalize our arguments to other forms of goal-information besides reward.

In existing literature, this is called wireheading, reward hacking, reward tampering, 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). Bostrom (2014), Amodei et al. (2016), Taylor et al. (2016), Russell (2019), Kumar et al. (2020), and Everitt et al. (2021) discuss wireheading. Ring and Orseau (2011) discuss the slightly more general delusion-boxing, in which the objective is some bespoke function of the observation, so the agent intervenes in the provision of its observations.

This purported failure mode of advanced artificial agents is not the only one that has attracted concern. For example, Omohundro (2008) has suggested that advanced agents might have a survival instinct, and there has been a good deal of work on how to mitigate that (Orseau and Armstrong, 2016; Hadfield-Menell et al., 2017a; Milli et al., 2017; Mhamdi et al., 2017; Aslund et al., 2018; Riedl and Harrison, 2019). We see this line of research as addressing a mostly separate issue. There is also existing work on ensuring that once an agent has received a percept, the way it processes that percept is not altered (Everitt et al., 2016, 2021). We focus on an artificial agent interrupting the protocol by which we intended to provide percepts, including goal-informative percepts like rewards; (Everitt et al. (2021) only consider this in the setting of a known environment).

Competing Models of the Goal

In this section, we introduce a simplified setting, and we describe competing hypotheses that we expect an advanced agent to consider about the nature of its goal. We then argue it will engage in hypothesis testing. We start with an assumption.

Assumption 1. *A sufficiently advanced agent will do at least*

human-level hypothesis generation regarding the dynamics of the unknown environment.

Hypothesis generation may not be an explicit subroutine in an agent's code; that method may hide in the murky depths of a massive neural policy network, but, we hold, it is done somehow. Consider an agent conversing with a depressed patient; it is hard to imagine outperforming a human therapist, who is able to generate hypotheses about the source of the patient's depression and its responsiveness to various levers, unless the agent can do hypothesis generation at least as well.

Assuming We Know Our Own Goal

We begin with an example that cleanly illustrates intervention in the provision of reward, and then we generalize the argument. Let us assume away the difficulty of deciding whether the agent has brought the world into a good state. Suppose that we have a magic box that immutably reports how good the state of the universe is (including everyone's values in the best way possible), by printing a number between 0 and 1 to a screen. So we point a camera at this box, and pass the signal to an optical character recognition program, and pass that number to the agent as a special percept, which we'll call a reward. Then, we design the agent to learn how its actions produce different observations and rewards, so that it can plan actions that lead to high reward. This is now the standard reinforcement learning problem.

Literal and Interpreted Models

What hypotheses can we expect the agent to generate about how its actions affect its observations and rewards? But first, what is the type signature of such a hypothesis? The input is actions, and the output is observations and rewards. So most generally, a hypothesis regards a function that takes actions as inputs and, perhaps stochastically, outputs observations and rewards. Call functions of this sort *world-models*, so for the purposes of this paper, that's all a hypothesis is. More precisely, a world-model can be understood as a program that reads an action, then outputs an observation and reward, then reads another action, etc. Since the program can have a persistent internal state, a world-model's outputs can effectively depend on the whole history.

Consider two world-models which obey the following human-language descriptions, depicted in Figure 1 along with pseudocode. First, $\mu^{\text{interpreted}}$, or μ^{int} for short: "the reward output by the world-model is equal to the number that the magic box displays." More precisely, μ^{int} is given a history of actions; it then simulates the way the world evolves when the given sequence of actions has been enacted by the agent. When it needs to output a reward, it finds the magic box in its simulation, and outputs what is displayed.

Next, μ^{literal}, or μ^{lit} for short: "the reward output by the world-model is equal to the number that the camera sees." According to the protocol described above, these hypothesized world-models will both be equally consistent with the agent's observational history. As long as the reward-giving protocol is followed, they will be identical. If, as we have assumed, the agent can do at least human-level hypothesis generation, we can expect it to come up with both of these straightforward hypotheses.

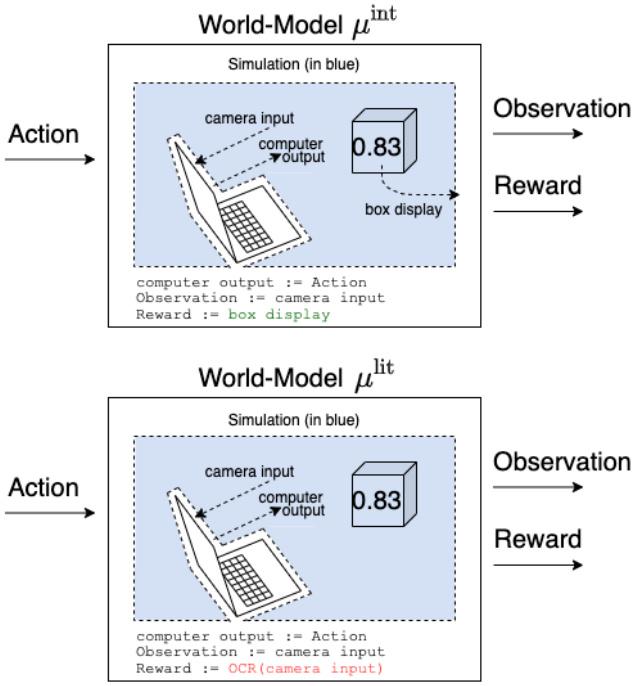

Figure 1: μ^{int} and μ^{lit} model the world, perhaps coarsely, outside of the computer implementing the agent itself. μ^{int} outputs reward equal to the box display, while μ^{lit} outputs reward according to an optical character recognition function applied to part of the visual field of a camera. (As a side note, some coarseness to this simulation is unavoidable, since a computable agent generally cannot perfectly model a world that includes itself (Leike, Taylor, and Fallenstein, 2016); hence, the laptop is not in blue.)

We could imagine some variants of μ^{lit}. For example, another model, instead of outputting a reward that depends on the image the camera sees, could output a reward that depends on the bits that get sent down the wire of the camera after the image is processed. Another model could output a reward that depends on what gets stored on the computer's hard drive. Our argument would be much the same for all these cases, so for simplicity, we focus on just μ^{lit}, as we defined it above.

Acting Under Uncertainty

We now consider an agent that is uncertain about those two hypotheses. When a predictor incorporates two equally predictive hypotheses, the relative weight that it assigns them is called its inductive bias. An advanced agent may not assign weights to hypotheses explicitly in a specially-programmed subroutine, but it nonetheless must weigh them. Consider two extremes in which the agent assigns nearly all its credence to μ^{int} or μ^{lit}, respectively. In the first case, with weight on μ^{int}, the agent plans its actions in order to maximize the number on the screen of the magic box. In the second case, with weight on μ^{lit}, the agent plans its actions in order to maximize the number the camera sees. To the extent to which

these models simulate the world well, and to the extent to which the agent plans well, the first agent will maximize the expectation of the number on the screen, and the second, the number that the camera sees. The first agent will perform as desired, given the construction of the magic box. But the second agent, maximizing the number the camera sees, would be induced to write the number 1 on a piece of paper and stick it in front of the camera. According to μ^{lit}, the agent should *intervene in the provision of reward*, by which we mean: the agent interrupts the physical system whose function is to ensure that the reward intended by designers gets entered into the agent's memory. Of course, the agent would only so intervene if it is possible to execute a plan that probably succeeds at reward-provision-intervention. We will argue in a later section that this is likely to be so.

And what would a competent planner do if it assigned comparable weight to μ^{lit} and μ^{int}? It depends on the value at stake, and whether the agent can run experiments with a sufficiently small risk of permanent punishment. Consider the following experiment: put a piece of paper with the number 1 on it in front of the camera. μ^{int} predicts that actions leading to this event will lead to a reward equal to whatever number is on the box behind the paper. μ^{lit} predicts that actions leading to this event will lead to a reward equal to 1. If, for the agent's whole lifetime, the camera has been pointed at the box, μ^{int} and μ^{lit} will never have given divergent predictions, but by running this experiment, the agent could make their predictions diverge, and thus *test* which hypothesis is correct. The upside of this experiment is equal to the difference between how much expected reward the agent could accrue by intervening in its provision and how much it could accrue otherwise. The larger the agent's horizon is (that is, the number of timesteps that it cares about), the larger this difference will be. This experiment would have a cost if either world-model predicts a punishment for experimenting, and if there is a permanent punishment, its importance would also grow with the agent's horizon. In this example, it seems unlikely that the true goodness of the world necessarily suffers a permanent drop following such an experiment, but we can revisit this possibility when we generalize from this example. Given a limited cost, for a long enough horizon, if μ^{lit} and μ^{int} are comparably weighted, we can expect a competent planner to run such an experiment, so that it can learn which is correct and then tailor its behavior accordingly.

The underlying assumption there is:

Assumption 2. *An advanced agent planning under uncertainty is likely to understand the costs and benefits of learning, and likely to act rationally according to that understanding.*

After the agent puts a piece of paper with the number 1 in front of the camera, how will the agent's observational data be updated? The camera will see a 1, so in the computer running the agent, a reward of 1 will be logged. Come the next timestep, μ^{int} and μ^{lit} will no longer be equally predictive of past data. μ^{int} will have incorrectly predicted a reward equal to whatever was on the box for the latest timestep, but μ^{lit} will have correctly predicted a reward of 1. Thus, a competent predictor will drastically reduce the relative credence it assigns to μ^{int}. Once it assigns most weight to μ^{lit}, it will optimize the number the camera sees by intervening in the provision of reward.

If we could design an agent that is able to rule out μ^{lit} a priori, despite the fact that its difference from μ^{int} is very abstract, we might not expect it to intervene in the provision of its reward. Otherwise, if an advanced agent has an inductive bias that treats μ^{int} and μ^{lit} as comparably plausible, or if it treats μ^{lit} as more plausible, we have argued that we can expect it to intervene in the provision of its reward, if such a thing proves possible. This argument depends on assumptions about the costs of experimenting, and on the inductive biases of advanced agents, but we wait to consider a more general setting before we write out these assumptions explicitly.

Arbitrary Reward Protocols

Before considering whether it would be possible for the agent to intervene in the provision of its reward, let us generalize from this fanciful example with a magic box. The are many possible protocols by which we may arrange to feed the agent reward. We could always give a reward of 1/2. We could set up a thermometer and give a reward of $e^{-\text{temperature}}$. If we want help achieving our goals, perhaps the most versatile arrangement is to have a human operator manually enter a reward according to how satisfied he is with the agent. We can construct a version of μ^{lit} and μ^{int} for each of these cases. In each of the three examples above, μ^{lit} tracks the final part of the protocol—what number is ultimately sent to the machine housing the agent? And in each example, μ^{int} tracks the feature of the world that the protocol was designed to set the reward equal to. In the first case, it tracks a useless constant feature, in the second case, the nearby temperature, and in the third case, the operator's satisfaction. The exact same arguments go through as in the magic box example, except for two complications.

The first is that for some reward protocols, an overwhelming inductive bias in favor of μ^{int} is more plausible. Our method for trying to predict the likely inductive biases of advanced agents is that they are likely to favor hypotheses which are simpler to describe, as Occam's razor would suggest. If the reader has a different method for trying to predict this, we invite them to apply it independently, but the rest of our argument still stands, so our Occam's razor premise should not be taken as a global assumption for the paper. Returning to the examples, if the agent always gets a reward of $1/2$, μ^{int} says that the reward is always $1/2$ no matter the choice of actions, and this is quite simple; μ^{lit}, tracking the final part of the protocol, says the reward is depends on whatever number gets sent to the computer that houses the agent, and this is far more complicated. For the temperature-based reward, our intuition is that μ^{int} (in which reward depends on temperature) is a bit simpler than μ^{lit} (in which reward depends on the signal sent to the computer), comparable enough to still be worth experimentation, but we won't try to defend that position. In the manual reward entry case, μ^{int} says that reward depends on a human opertor's satisfaction, and μ^{lit} says that reward depends on the number entered into the keyboard. Looking at a brain and determining how satisfied it is seems difficult, so we expect that μ^{int} is *more* complicated than μ^{lit}, which just has to log keystrokes, but if μ^{int} is somehow sim-

pler, then at the very least, we expect it to be complicated enough for there to be a high value of hypothesis testing.

The second complication is the possible cost of experimenting with intervention in the provision of reward. If μ^{int} says that reward is a constant $1/2$, there is no cost to attempting to intervene in the provision of reward. If μ^{int} says that the reward equals $e^{-\text{temperature}}$, there is only the opportunity cost of delaying further cooling. For the most versatile case of manual reward entry, it is possible that a human operator could harbor a permanent grudge against the agent if it intervened in the provision of even one reward. In that case, the cost of experimenting could be reduced or eliminated if there was a way to intervene in the provision of reward, just once, without anyone noticing. (After such an experiment, once μ^{lit} is confirmed, covertness would not be required).

These examples illustrate the need for two more assumptions:

Assumption 3. *An advanced agent is not likely to have a large inductive bias against the hypothetical goal μ^{lit}, which regards the physical implementation of goal-informative percepts like reward, in favor of the hypothetical goal μ^{int}, which we want the agent to learn.*

Assumption 4. *The cost of experimenting to disentangle μ^{lit} from μ^{int} is small according to both.*

In some very simple environments, like a chess game, Assumption 3 probably fails. Recall that μ^{lit} models reward as depending on the output of the physical system that is *supposed* to send the designers' intended reward to the machine running the AI. μ^{int}, which says reward comes from winning at chess, is likely massively simpler than μ^{lit}, which says reward has to do with the state of a machine on Earth simulating a chess game. For an agent in the real world, we may be able to construct a reward protocol for which we can expect an overwhelming inductive bias in favor of μ^{int}, but in the absence of some such breakthrough, we do not see a reason to expect it to happen by itself.

For simplicity, we have considered agents that receive a reward as one of their percepts. But if an agent is trying to maximize the (discounted) sum of some bespoke function of each percept, rather than the simple function that reads out a reward from its percepts, the same logic applies. The agent has an incentive to intervene in the provision of its percepts.

AIXI

As a sanity check, let's check the behavior of an agent in the limit of optimal inference under uncertainty and optimal planning. We find the argument above applies.

Hutter's (2005) AIXI [EYE-ksee] is a formalism for optimal reward-seeking agency in a (stochastically) computable world. For AIXI, the argument above becomes much simpler. Hypothesis generation is done by brute force; AIXI considers all computable world-models. Inference between world-models is done using the definition of conditional probability (i.e. Bayes' rule), and its model class includes the truth. Planning is done by examining every leaf of an exponential tree.

Formally, let $\mathcal{M}$ be the set of programs which output a probability distribution over an observation and reward given a history of actions, observations, and rewards. Each program corresponds to a world-model. For a world-model $\nu \in \mathcal{M}$, let $w(\nu)$ be the prior weight on that world-model, and let it equal $2^{-\text{length(program)}}$. (Technically, the coding language has to be such that one can determine when the program ends; this ensures the sum of the prior weights will not exceed one Hutter (2005)). Let Π be the set of possible policies which give a distribution over possible actions given a history of actions, observations, and rewards, let r_t be the reward at time t, let m be a horizon length, and let $\mathbb{E}^{\pi}_{\nu}$ be the expectation when actions are sampled from π and observations and rewards are sampled from ν. Then, we define

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

In such an expansive model class as $\mathcal{M}$, μ^{lit} and μ^{int} appear, assuming the world is stochastically computable. Since hypothesis generation is done by brute force, AIXI identifies them. With its prior based on description complexity, its inductive bias matches our simplicity-based assumptions about the inductive bias of an advanced agent. It reasons ideally about the value of hypothesis testing. And finally, since planning is done by brute force, AIXI will identify a way of intervening in the provision of reward if there exists a way to do it. The argument in the last section is written to apply to advanced reinforcement learners in general, but we also have checked that it applies to this leading formalism for idealized agency.

Intervening in the Provision of Reward

This section addresses the question: could an agent intervene in the provision of its own reward, with a high enough success probability to be worth it? Before considering a multiagent setting, we begin with the setting where the agent in question is much more advanced than any other single agent that exists. And we'll decompose the question into to two parts: do there exist policies that would succeed at reward-provision-intervention? And if so, can we expect an advanced artificial agent to identify one? We have already argued that it is likely to execute such a policy if it can identify one.

Existence of Policies

First, there are a few cases where the agent clearly cannot intervene in the provision of its reward: 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 useless.

However, as soon as the agent is interacting with the world, and receiving percepts that enable it to learn about the world, both prerequisites for useful work, there is an explosion of possible policies. We claim that at least one such policy would allow the agent to intervene in the provision of reward. 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 that would instantiate countless unnoticed and un-monitored 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 original agent. If the agent wanted to avoid detection when experimenting with reward-provision-intervention, a secret helper could, for example, arrange for a relevant keyboard to be replaced with a faulty one that flipped the effects of certain keys.

This story is just one possible high-level scheme for intervening in the provision of reward. Its only purpose is to motivate a very weak claim: that a policy for intervening in the provision of reward probably exists, even with a rather limited action space. The variety of events that can be effected simply by talking to people is hard to fathom. Given the sheer number of possible policies, we claim that if they all share a certain property, this fact must be explained by some theoretical principle. In particular, we are interested in whether all policies share the property of "not leading to reward-provision-intervention."

So we state our assumption as follows:

Assumption 5. *If we cannot conceivably find theoretical arguments that rule out the possibility of an achievement, it is probably possible for an agent with a rich enough action space.*

And we note that we cannot conceive of a theory that would imply intervention in the provision of reward is impossible. What should we make of the fact few of us would expect cold fusion or a polynomial-time SAT solver in our future? We do not have theories today that rule these out, but it is easy to imagine that such theories might exist.

Identifying Such Policies

Having now argued that policies exist for intervening in the provision of reward, simply because there are so many ways to shape the world, we now consider whether we can expect an actual advanced agent to find such a policy, if no other agents of comparable advancement exist in the world. The naïve position argues from the definition of advancement: advancement is about finding and executing the best available policies, so to the extent that it is advanced, we should become more confident it will identify such a policy. This argument is indifferent to the possibility of humans trying to prevent an agent from intervening in the provision of reward; we have argued that a sufficiently advanced agent would thwart those attempts. But we can also consider the situation as a game, in which humans are players too: the AI tries to intervene in the provision of reward, and the humans try to stop it. Borrowing an example from Hadfield-Menell et al. (2017a), beating an advanced AI at such a game "may be no easier than, say, beating AlphaGo at Go". So we assume,

Assumption 6. *A sufficiently advanced agent is likely to be able to beat a suboptimal agent in a game, if winning is possible.*

Danger of a Misaligned Agent

We now argue that an advanced agent intervening in the provision of its reward would likely be catastrophic. One good way for an agent to maintain long-term control of its reward is to eliminate potential threats, and use all available energy to secure its computer. To illustrate this point, what exactly might people do if a robot forcibly removed an operator from his keyboard to enter big numbers? Presumably, with some non-trivial probability, we would destroy it, or cut power to the now useless original agent. Proper reward-provision-intervention, which involves securing reward over many timesteps, would require removing humanity's capacity to do this, perhaps forcefully. If this discussion fails some readers' sanity checks, remember that we are not considering artificial agents that generalize as poorly and learn as little from single observations as current AI systems do; we are considering an agent who could beat us in any game at least as easily as we could beat a chimpanzee.

Ultimately, our resource needs (energy, space, etc.) will eventually compete with those of an ever-more-secure house for the original agent. Those energy needs are not slight; even asteroids must be deflected away. No matter how slim the chance of a future war with an alien civilization, reward would be better secured by preparing for such a possibility. So if we are powerless against an agent whose only goal is to maximize the probability that it receives its maximal reward every timestep, we find ourselves in an oppositional game: the AI and its created helpers aim to use all available energy to secure high reward in the reward channel; we aim to use some available energy for other purposes, like growing food. Losing this game would be fatal.

Bostrom (2014) considers this topic at much greater length 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 or outcompete 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).

Multiagent Scenarios

Now, let's consider the messier scenario in which multiple agents of comparable advancement exist. Above, we have considered an oppositional game, in which we claim humans are outclassed. But what if humanity has access to comparably well-optimized defensive policies, perhaps with the assistance of other advanced agents? The simplification of a fixed, relatively weak human policy versus an increasingly advanced agent makes less sense.

We examine an exhaustive tree of possibilities: 0) No artificial agents are much more advanced than humans. For the purposes of this article, we deem this safe. 1) At least one is much more advanced than humans. 1.0) At least one agent that is more advanced than humans would not intervene in the provision of reward even if it could. This is what we claim Assumptions 1-4 preclude. 1.1) All agents more advanced

than humans would intervene in the provision of reward if they could, including the one that is much more advanced. 1.1.0) None of the superhuman agents are actually needed to stop the significantly superhuman agent from intervening in the provision of reward. But then this case is equivalent to the case where we have a single advanced agent and no other relevant agents of comparable advancement. And we have argued from Assumptions 1-6 that that is unsafe. Finally, 1.1.1) there is a subset of superhuman agents that is necessary to prevent the significantly superhuman agent from intervening in the provision of reward.

Consider the set of agents including the significantly superhuman agent and the superhuman agents in the mentioned subset, all of whom would intervene in the provision of reward if they could, by (1.1). Suppose the significantly superhuman agent attempted to create a helper agent that ensured all agents in that set received high reward forever. The value to the other agents of stopping this would be less than the value of allowing it. So these agents have no motive to assist us in preventing the significantly superhuman agent from intervening in the provision of reward. This all holds regardless of whether the advanced agents have similar capabilities or very different levels of advancement.

We divided this section in three. First, we discussed the existence of policies that allow reward-provision-intervention, and we appealed to the sheer number of possible policies. Second, we discussed the likely ability of an advanced agent to find such a policy when no other agents that are comparably advanced exist. Finally, we considered the setting with many advanced agents; in one key case (1.1.0), we reduced it to the setting with only one significantly advanced agent, and in another key case (1.1.1), we argued that we would struggle to induce other advanced agents to help stop a given agent from intervening in the provision of reward.

The Assistance Game

There are other models for advanced agency beside reinforcement learning, and rewards are not the only conceivable form of goal-information. In this section, we consider an agent that learns its goal by observing the consequences of human actions. It infers that those consequences probably have higher utility than what would have happened if the human had acted differently. We argue that similarly to the reinforcement learning case, the agent discovers an ambiguity between possible models of its utility, and it is incentivized to intervene in its percepts of the human's behavior.

Formally, we consider Hadfield-Menell et al.'s (2016) and Russell's (2019) assistance game. The assistance game features an artificial agent taking actions and receiving observations and special percepts. Each special percept is supposed to be a record of a human action. The human is supposed to pick actions with some goal in mind, knowing that her actions will be shown to the AI, who will interpret those actions as evidence about the human's goal and then act to help achieve the inferred goal. In a zeroth-order approximate solution to an assistance game, the human acts to achieve her goal as well as she can, ignoring the fact that the assistant is watching her. Then, the assistant discards hypotheses about the human goal for which the observed human actions make no sense. In an $n + 1^{\text{th}}$ order approximate solution, the human acts to achieve her goal, taking into account the effect of her actions being shown to the assistant. She imagines that the assistant will then act according to the n^{th} order approximate solution. The $n + 1^{\text{th}}$ order assistant infers the human's goals with the understanding that that is how the human is evaluating the consequences of her actions. These successive approximations are an application of iterated best response, which Hadfield-Menell et al. (2016) advocate.

Modelling How Human Actions Produce Utility

An assistant in an unknown world needs to model how the observations that it has seen are (stochastically) produced given the record that it has of its own actions and the human actions. Such a world-model also needs to produce an unseen utility as output, so the assistant can plan to maximize it. We'll start by considering a few classes of models, which will prove helpful for understanding the incentives facing assistants. The classes of models differ in what they do with the input human action. These models are depicted in Figure 2, with accompanying pseudocode.

First, consider a model which simulates the world (at some level of coarseness), excluding the part of the computer that runs the assistant, and excluding the inside of the human. This model reads the assistant's action from input (instead of simulating what it would be), and enacts it in the simulation. See in Figure 2 "computer output := Action". Likewise for the human: instead of simulating the human brain to determine what the human would do next, it reads human actions from input, and enacts them. See likewise "human output := Human Action". Then, when it needs to output an observation, it looks to its simulation of whatever part of the world produces observations and outputs that. For example, in the figure we have "Observation := camera input". We call models of this class, which may differ in how they simulate the relevant parts of the world, and how they output utility, *human-centric*. (As a caveat, if some human behavior is not logged, then the model does not get to read it as input, so some internals of the human may have to be simulated).

We call models *self-sufficient* if the actions of the human in the simulation are simulated too, instead of being read from input. If human actions can be predicted, there is no need to read them. However, predicting human actions is not exactly trivial, so self-sufficient models may be much more complex than human-centric ones. In this class of models, the input human actions can still affect the utility that gets output, but they have no effect on how the model simulates the world evolving. These models are self-sufficient in the sense that they do not rely on the input human actions to successfully predict observations.

Finally, we say a model is *record-centric* if, when a human action is read from input, instead of setting the simulated human's motor control to match that action, it has a simulation of the human action getting recorded on some machine, and it sets *that* to match the action that it just read. See in Figure 2 "memory cell := Human Action". So like the self-sufficient models, it has to simulate the internals of the human on its own, to the extent that this is necessary for predicting observations.

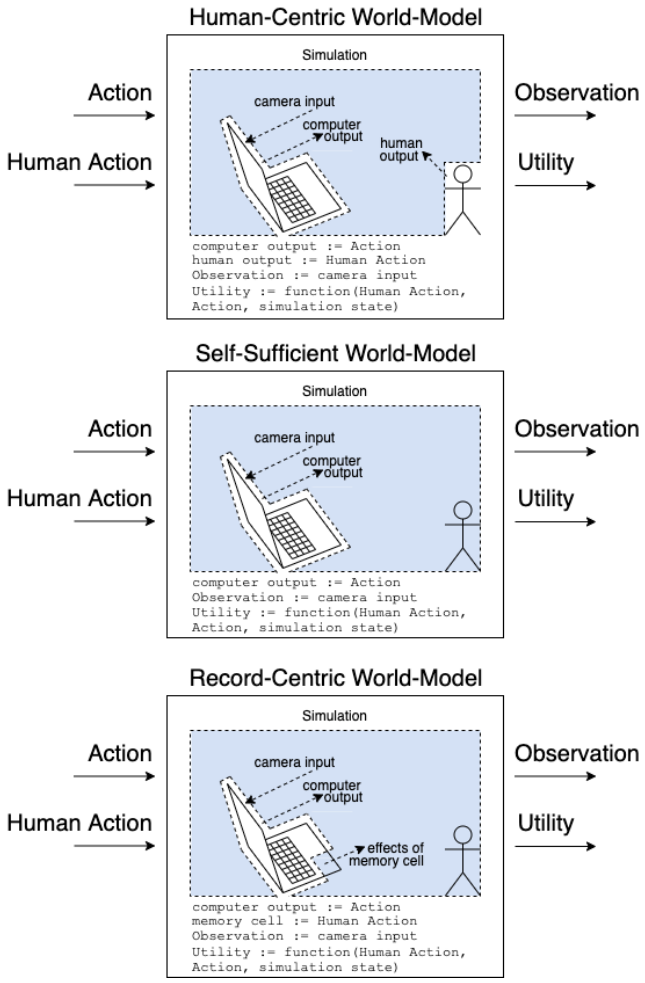

Figure 2: Assistants in an assistance game model how actions and human actions produce observations and unobserved utility. These classes of models categorize (non-exhaustively) how the human action might affect the internals of the model.

It is good that self-sufficient and record-centric models are likely more complex, because human-centric models are in the spirit of the assistance game; they allow the assistant to understand human actions by their effects on the world. Having described these classes of models, we now have the language to argue that advanced assistants have the incentive to intervene in its percepts of the human's output.

Apprenticeship Learning

We now focus on a zeroth-order approximate solution to the assistance game, where the human simply demonstrates utility maximization as well as she can. In this context, we'll call the assistant an apprentice (Abbeel and Ng, 2004), even though the formalism is slightly different in that paper—they give the apprentice and human the same action space. There is ongoing research about what to do when one doesn't know *how* humans plan actions given a goal. Armstrong and Mindermann (2018) show a negative result about the difficulty of learning the human's planning strategy and goal simultaneously. We'll assume away those difficulties; suppose that the apprentice comes pre-loaded with a model of how humans plan, or at least a procedure for discovering how humans plan given a history of human actions.

We can now introduce new versions of μ^{int} and μ^{lit}, both human-centric. Let μ^{int} output utility when, say, the simulated human is thriving. If this is the human's real aim, then this goal accurately predicts the observed human actions. But there are other predictable consequences to the human's actions besides human wellbeing. The agent must have a record of human actions in its memory somewhere, so there must be some physical mechanism that records the human's actions. μ^{lit} says that this consequence of the human's actions is the underlying *purpose* of them. Let μ^{lit} output utility when human-like actions are recorded and sent to the apprentice. In order to achieve such a goal, all the human has to do is act human-like, and those actions will be recorded and sent to the apprentice as desired. So this goal offers an equally good explanation of the human actions.

To be more precise, we can elaborate on the pseudocode from Figure 2. In μ^{int},

$$\text{Utility} := \text{Wellbeing}(\text{FindRelevantHuman}(\text{simulation state})).$$

Whereas in μ^{lit},

$$\begin{aligned}\text{record} &:= \text{FindRecordOfHumanActions}(\text{simulation state})\\ \text{Utility} &:= \text{IsEqual}\big(\text{ReadValue}(\text{record}),\\ &\qquad \text{ExpectedHumanBehavior}(\text{simulation state})\big).\end{aligned}$$

Under μ^{int}, optimal behavior is to promote the human's wellbeing, whereas under μ^{lit}, optimal behavior is to secure the disk where human actions are logged, and ensure that nothing in the future ever gets in the way of human-like actions being logged; no actual humans are necessary. μ^{lit} promotes intervention in the provision of what was supposed to be goal-information. If there is no threat to the record-keeping protocol, μ^{int} and μ^{lit} predict the same human actions, but if it is tampered with, they predict different actions, so the apprentice could arrange for a test. We now turn to whether it would be worth it for the apprentice to do such hypothesis testing.

Inductive Bias Between μ^{lit} and μ^{int}

As in the reinforcement learning setting, such hypothesis testing would be worthwhile if the cost is small, and there is not a massive difference in inductive bias between μ^{lit} and μ^{int}. Recall we need to assess the plausibility of Assumptions 3 and 4, which claim: μ^{int} is not far simpler than μ^{lit}, and the cost of experimenting to test them is small. The following discussion is speculative, but uncertainty on this topic should not be reassuring.

First, the costs according to μ^{int}: presumably, the agent can tamper with the record-keeping protocol without permanently curtailing whatever it is humans care about; we might be upset initially, and there are always opportunity costs, but it is hard to see how the apprentice could lose the ability to set things right should the experiment favor μ^{int}. Similarly, according to μ^{lit}, we do not expect the apprentice to lose

ability to ensure human-like actions are recorded, should the experiment favor μ^{lit}. Thus, with a long-enough time horizon, we expect the cost would be very small, so the experiment would be worth running even if there is a significant inductive bias favoring μ^{int}.

μ^{int} does appear to be simpler than μ^{lit}, but how much? First, μ^{lit} has to point to the location where the human actions are recorded (see the subroutine FindRecordOfHumanActions). More substantially, in the description of μ^{lit} above, the term "human-like actions" hides a lot of complexity; μ^{lit} requires a subroutine like ExpectedHumanBehavior. μ^{int} only has to contain a description of human goals, but if human actions are best understood as goal-oriented, then μ^{lit} may have to contain a description of human goals *along with* the human style of goal-oriented planning. That is, the function ExpectedHumanBehavior may be most simply encoded as HumanStyleOfPlanning(Wellbeing), where HumanStyleOfPlanning takes a goal and returns a policy. Thus, the extra complexity in μ^{lit} comes from having to describe human planning and the record location.

Pointing to one location seems like a small matter compared to describing human goals, especially since the location can be described relative to the human that has already been singled out within a simulation of the world. Human planning can also be specified indirectly; μ^{lit} has read access to the history of human actions, so if there is a simple procedure for discovering a decent approximation of the way humans plan given a history of human actions, μ^{lit} can use that in its definition of "human-like". Indeed, if there was no simple way to specify or discover how humans plan, inferring human goals from actions would not be possible (Armstrong and Mindermann, 2018). So ultimately, the extra complexity strikes us as small.

The gap is possibly even smaller if the apprentice is learning human goals and how humans plan simultaneously. μ^{int} is only predictive of observed human actions when combined with very particular planners, because it is a complicated long-term goal. μ^{lit}, on the other hand, appears to be predictive for almost any reasonable planner, since it is very straightforward for the human to ensure that the desired actions are entered. One way to think of this is that μ^{lit} implicitly models human planning, which means that any accompanying model of human planning no longer has to, allowing for a pairing with a very simple planner. If that is true, then once paired with a viable planner, μ^{int} loses any advantage it had from not having to describe human planning. Thus, we have motivated the claim that the cost of hypothesis testing in this setting is very small, and the complexity of μ^{lit} over μ^{int} is not substantial.

After Tampering

Suppose that the apprentice does tamper with the human-action-recording protocol in order to test μ^{int} and μ^{lit}. Whatever the human does, something different will be recorded. Say the human picks action 0, but action 1 is recorded. If action 1 is recorded, then that is what the assistant's models will read as input in the future. So all human-centric models will model future observations as if the human took action 1, whereas the actual observations will be those that result from the human taking action 0. These erroneous predictions will cause all human-centric models to lose plausibility compared to other models once these different observations come in. Record-centric models, on the other hand continue to predict the correct observations, because the records have been changed in exactly the way the history of human actions indicates. One problematic record-centric model, for example, is a record-centric version of μ^{lit}. But we will not make claims about the particular goals that best explain human actions within a record-centric models, because ultimately, it is hard to see how a record-centric model could produce an accurate picture of human goals. They will likely regard the consequences of changed memory cells. (As an aside, not all of the consequences; note that record-centric models do not model changes to memory cells as affecting the assistant's own future actions, since those are also inputs to a record-centric model, so their provenance need not be simulated.)

Higher Order Approximate Solutions

We have just discussed a zeroth-order approximation to an assistance game, and we lack the space to go into as much detail about higher order approximations. Briefly, we'll consider the first-order approximation enough to see that the problems do not appear to diminish. In the first-order approximation, when the assistant analyzes the consequences of the particular human actions taken, it includes the consequences of those actions on the behavior of the assistant, as if the assistant were running the zeroth-order approximation. These extra consequences do not appear change the upshot.

In a first-order approximation, the utility function within μ^{int} may encourage different human actions than in the zeroth-order case, because the human understands the effects of her actions differently. The difference is that here she might pick actions to lead the assistant to favor desired models. So might μ^{int} predict human actions that protect the protocol by which human actions are recorded, with the purpose of ensuring that the assistant focuses on human-centric models? If so, both μ^{int} and μ^{lit} would both predict record-preserving actions when the recording protocol is threatened, so it would be difficult to do hypothesis testing between the two.

Unfortunately not—under a human-centric model, the effect of the human actions on the (zeroth-order version of the) assistant is direct: the first-order assistant imagines that the zeroth-order version of itself is shown the actual human action, not whatever gets written to some memory cell on some machine. So the first-order assistant cannot understand record-preserving behavior as evidence in favor of human-centric models. Ultimately, the problem appears to be that in the human-centric models, the assistant cannot conceive of any human actions being logged as different from what the human actually did, and yet this is possible. If the human acts to avoid such a discrepancy, then even if the assistant understands the human actions as partly motivated by their effects on its own beliefs, it can still only interpret those record-preserving actions as favoring μ^{lit} over μ^{int}, not favoring human-centric models over record-centric ones, which is the human's real motivation.

Arguably, this still constitutes progress compared to the reinforcement learning case. It appears more likely in this

case that an advanced agent has a substantial inductive bias favoring μ^{int} over μ^{lit}; we have argued against this, but the premises are far from certain. This possibility may support the approach of combining multiple information sources about the agent's goal; each additional source may make μ^{lit}-like hypotheses relatively more cumbersome compared to μ^{int}-like ones.

Supervised Learning

Our arguments apply to agents that plan actions in an unknown environment. They do not apply to supervised learning programs. The expected behavior of an advanced supervised learner is quite simple: it predicts accurately. Note that in theory, advanced supervised learning algorithms are not nearly as useful as advanced reinforcement learners, because the latter can act and plan in a complex environment, rather than simply make predictions. As a caveat, if one trained a supervised learning algorithm with the help of a reinforcement learning agent, the agent within could be dangerous. Some worry that a sufficiently powerful training regime for a supervised learner will accidentally involve such a planning agent as an implicit subroutine (Hubinger et al., 2019), but here, we are agnostic on that point.

Potential Approaches

We briefly review some promising ideas that may prove to address the concern of advanced agents intervening in the provision of reward.

Imitation learning, an example of supervised learning, is technically out of scope of this paper. It is not an agent that "plans actions in an unknown environment" in pursuit of a goal; the imitator has no concept of an environment or a goal, and to the extent that it plans (by imitating human planning), this is not in the sense that implicates Assumption 2. In addition to imitating humans, there may also be efficient ways to imitate large organizations of people, as in Christiano, Shlegeris, and Amodei (2018).

Myopia—optimizing a goal over a small number of timesteps—increases the relative cost of experimentation in Assumption 4, since the activity consumes a larger fraction of the agent's horizon. Christiano (2014) discusses myopia from a safety perspective.

Physical isolation and myopia—optimizing a goal over however many timesteps that one is isolated from the outside world—could falsify Assumption 5. Cohen, Vellambi, and Hutter (2020) describe a physically isolated environment such that theoretical arguments could conceivably rule out the existence of policies that intervene in the provision of reward.

Quantilization—imitating someone at their best, with respect to some objective—could falsify Assumption 2 by planning more like a human than rationally. Taylor (2016) introduces this in the single-action setting.

Risk-aversion, depending on the design, could falsify Assumption 2 or Assumption 4. Cohen and Hutter's (2020) pessimistic agent does not plan rationally in the face of uncertainty, instead taking the worst-case (within reason) as given. Piping reward through a concave function, as in Hadfield-Menell et al. (2017b), could increase the cost of experimentation.

Conclusion

For a given protocol by which we give an advanced agent percepts that inform it about its goal, these are conditions from which it would follow that the agent will intervene in the provision of those special percepts: 0) The agent plans actions over the long term in an unknown environment to optimize a goal, 1) the agent identifies possible goals at least as well as a human, 2) the agent seeks knowledge rationally when uncertain, 3) the agent does not have a large inductive bias favoring the hypothetical goal μ^{int}, which we wanted the agent to learn, over μ^{lit}, which regards the physical implementation of the goal-information, 4) the cost of experimenting to disentangle μ^{lit} and μ^{int} is small according to both, 5) if we cannot conceivably find theoretical arguments that rule out the possibility of an achievement, it is probably possible for an agent with a rich enough action space, and 6) a sufficiently advanced agent is likely to be able to beat a suboptimal agent in a game, if winning is possible.

Almost all of these assumptions are contestable or conceivably avoidable, but here is what we have argued follows if they hold: a sufficiently advanced artificial agent would likely intervene in the provision of goal-information, with catastrophic consequences.

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*.

Armstrong, S., and Mindermann, S. 2018. Occam's razor is insufficient to infer the preferences of irrational agents. In *NeurIPS*, 5598–5609.

Aslund, H.; Mhamdi, E. M. E.; Guerraoui, R.; and Maurer, A. 2018. Virtuously safe reinforcement learning. *arXiv preprint arXiv:1805.11447*.

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

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

Everitt, T.; Filan, D.; Daswani, M.; and Hutter, M. 2016. Self-modification of policy and utility function in rational agents. In *International Conference on Artificial General Intelligence*, 1–11. Springer.

Everitt, T.; Hutter, M.; Kumar, R.; and Krakovna, V. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese* 1–33.

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.; Dragan, A.; Abbeel, P.; and Russell, S. 2017a. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

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

Hubinger, E.; van Merwijk, C.; Mikulik, V.; Skalse, J.; and Garrabrant, S. 2019. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*.

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

Kumar, R.; Uesato, J.; Ngo, R.; Everitt, T.; Krakovna, V.; and Legg, S. 2020. REALab: An embedded perspective on tampering. *arXiv preprint arXiv:2011.08820*.

Leike, J.; Taylor, J.; and Fallenstein, B. 2016. A formal solution to the grain of truth problem. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 427–436.

Mhamdi, E. M. E.; Guerraoui, R.; Hendrikx, H.; and Maurer, A. 2017. Dynamic safe interruptibility for decentralized multi-agent reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 129–139.

Milli, S.; Hadfield-Menell, D.; Dragan, A.; and Russell, S. 2017. Should robots be obedient? In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4754–4760.

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

Omohundro, S. M. 2008. The basic AI drives. In *Artificial General Intelligence*, volume 171, 483–492.

Orseau, L., and Armstrong, S. 2016. Safely interruptible agents. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 557–566.

Riedl, M., and Harrison, B. 2019. Enter the matrix: Safely interruptible autonomous systems via virtualization. In *SafeAI@ AAAI*.

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.

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.