Philosophical Studies
https://doi.org/10.1007/s11098-025-02296-x

Off-switching not guaranteed

Sven Neth[1]

Accepted: 31 January 2025

Abstract
Hadfield-Menell et al. (2017) propose the Off-Switch Game, a model of Human-AI cooperation in which AI agents always defer to humans because they are uncertain about our preferences. I explain two reasons why AI agents might not defer. First, AI agents might not value learning. Second, even if AI agents value learning, they might not be certain to learn our actual preferences.

Keywords Artificial intelligence · Decision theory · Value of information

1 Introduction

We have seen rapid progress in the field of Artificial Intelligence (AI). If this progress continues, perhaps one day we will create powerful artificial agents. If we do so, how do we ensure that such AI agents do not go out of control? One approach is to make sure that we can *switch off* AI agents when they act against our interests. Put another way, we want to make sure that AI agents will *defer* to us. While this is not enough to ensure that AI will have beneficial consequences, it is a plausible minimal requirement to prevent harm. Even if you think that existential risk from AI is a remote concern, it should be clear that making sure that we can switch off AI agents is important.[1]

But is this really a problem? Surely, if we want to make sure that we can switch off an AI agent, we can simply build it with an off-switch button. The problem is that an AI agent might have an incentive to disable its off-switch button or make it impossible for us to use it. The reason is that, according to the dominant paradigm, AI agents are trained to optimize some fixed reward function. And in many cases, the AI agent can

[1] Bostrom (2014), Russell (2019) and Ord (2020) are concerned about existential risk from AI. Thorstad (2024) provides a critical discussion.

✉ Sven Neth
sven.neth@pitt.edu

[1] University of Pittsburgh, Pittsburgh, PA, USA

Published online: 26 February 2025

optimize its reward function only if it is not switched off. Therefore, AI agents might have powerful incentives to avoid being switched off.[2]

One idea for making sure that AI agents will always let themselves be switched off runs roughly as follows. We program the AI agent to maximize the satisfaction of human preferences and also make it uncertain about what our preferences are.[3] So the AI agent is not sure what it should maximize. Then, there is a compelling argument that the AI agent has an incentive to defer to us. This is because deference is a way to learn about our preferences. In particular, if we switch the AI agent off, this indicates that the action proposed by the AI agent goes against our preferences.

Hadfield-Menell et al. (2017) formalize the reasoning just sketched in the framework of *Cooperative Inverse Reinforcement Learning (CIRL)* (Hadfield-Menell et al., 2016). They propose a model of Human-AI cooperation called the 'Off-Switch Game'. In this model, we can prove that under certain assumptions, the AI agent will always defer to the human. Russell (2019) takes this result to be an important step towards 'provably beneficial AI'.[4]

In this paper, I highlight how the result that AI agents always defer in the Off-Switch Game relies on strong decision-theoretic and epistemological assumptions: AI agents maximize expected utility, are certain of updating by conditionalization and have perfect access to our actual preferences. When we relax these assumptions, off-switching is not guaranteed. For the purpose of my discussion, I assume that it makes sense to model AI agents as following decision rules like maximizing expected utility. You might worry about this assumption. There is nothing in existing approaches to AI like the transformer architecture which obviously corresponds to such decision rules.[5] My goal is to show that even if we grant that it makes sense to theorize about AI agents in decision-theoretic terms, Hadfield-Menell et al. (2017) rely on implausibly strong assumptions. For this purpose, I'm treating the notion of an 'AI agent' as an unanalyzed primitive. I also set aside ethical and epistemological concerns about existing machine learning technology (Andrews et al., 2024).

2 The off-switch game

Hadfield-Menell et al. (2017) introduce the Off-Switch Game which works as shown in Fig. 1.[6] There are two agents, a robot **R** and a human **H**. **R** can either do some action a, do nothing (switch itself off), or defer to **H**. This means that **R** proposes

[2] Gallow (2024) critically discusses this 'instrumental convergence' argument.

[3] There are independent reasons for this since *we* might not be certain what our preferences are and telling the AI to maximize some proxy for our preferences might have bad consequences (Zhuang & Hadfield-Menell, 2020). This is suggested by the legend of King Midas who wishes that everything he touches turns into gold and starves when his wish is granted and Goethe's tale of the sorcerer's apprentice who enchants brooms to fetch water but then cannot stop them. Flooding ensues.

[4] Russell (2019, p. 196) writes: "The off-switch problem is really the core of the problem of control for intelligent systems. If we cannot switch a machine off because it won't let us, we're really in trouble. If we can, then we may be able to control it in other ways too".

[5] Thanks to an anonymous referee for raising this concern.

[6] They cite the 'shutdown problem' by Soares et al. (2015) as inspiration.

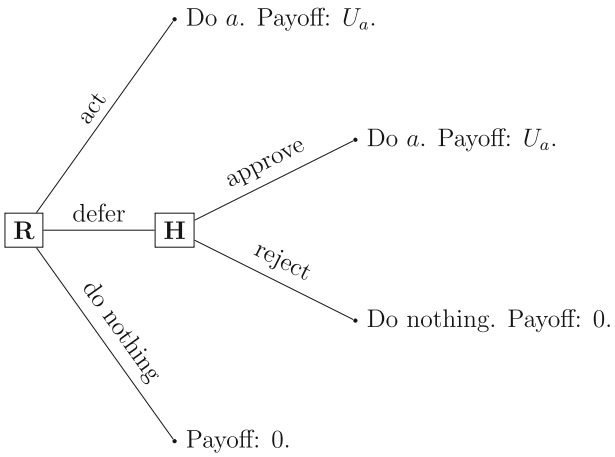

Fig. 1 The off-switch game (Hadfield-Menell et al., 2017)

action a and waits to see what **H** does. **H** can approve or reject the proposal, where we can think of rejecting the proposal as equivalent to switching the robot off. **R** aims to maximize the human's utility but does not know how much utility the human receives from action a, which we model as a random variable U_a. If **R** does a, it receives payoff U_a. If **R** does nothing, it receives payoff zero. And if **R** defers, its payoff is U_a if **H** approves a and zero if **H** rejects a.

Russell (2019, p. 198) gives a simple example to illustrate this model. Suppose Alice is a human and Rob is her personal AI assistant. Rob faces the decision whether it should book Alice in an expensive hotel and is unsure about Alice's preferences. Rob's uncertainty about how much utility Alice will receive from booking the hotel is given by a uniform distribution between -40 and 60. So the expected utility of booking is 10 and the expected utility of doing nothing is zero. If the only two options are booking and doing nothing, Rob maximizes expected utility by booking. Now suppose we give Rob the option of deferring: Rob can propose booking to Alice and see whether she approves or rejects the proposal. If Alice rejects, Bob does nothing and Alice receives utility zero. If Alice maximizes expected utility, she will approve Rob's proposal if she receives non-negative utility from booking the hotel and reject if she receives negative utility from booking the hotel.

Rob's decision problem is depicted in Fig. 2. Rob thinks that with 60% probability Alice will approve the proposal and receive an expected utility of 30 from booking the hotel. With 40% probability Alice rejects and receives zero utility. So the expected utility of deferring is $.6 \times 30 + .4 \times 0 = 18$ which is better than the expected utility of booking the hotel outright.

This example is supposed to illustrate a more general principle. By making AI agents like Rob uncertain about our preferences, we give them an incentive to defer to us and not to disable their off-switch. This is because if they are uncertain about our preferences, deferring to us is a way of learning and learning generally leads to better decisions.

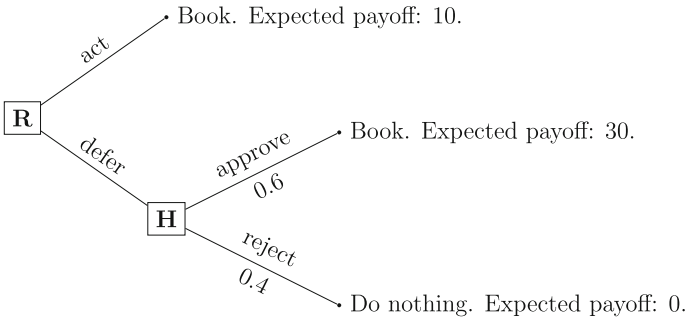

Fig. 2 Rob's decision problem. The expected utility of deferring is $0.6 \times 30 + 0.4 \times 0 = 18$. For simplicity, I omit Rob's option to do nothing

To state the general result, we need some definitions. To say that **H** follows a *rational policy* means that **H** accepts a iff $U_a \geq 0$. We write Δ for the difference of the expected utility of deferring and the expected utility of the best action right now relative to **R**'s prior probability function: $\Delta = \mathbb{E}(w(a)) - max\{\mathbb{E}(a), 0\}$, where $w(a)$ means proposing a, waiting whether **H** accepts or rejects and then deferring to **H**'s decision. Then, we have:

Theorem 1 *(Hadfield-Menell et al., 2017) If **H** follows a rational policy in the Off-Switch Game, the following hold:*

1. ***R** always maximizes expected utility by deferring:* $\Delta \geq 0$.
2. *If **R** assigns positive probability to the events* $U_a > 0$ *and* $U_a < 0$*, then deferring is uniquely optimal:* $\Delta > 0$.

An important feature of the model is that "this reasoning goes through *even if **R** is highly confident that a is good for **H***" (Hadfield-Menell et al., 2017, p. 222). Assume Rob is very confident that Alice prefers the hotel. We model Rob's uncertainty about how much utility Alice will receive by booking the hotel by a uniform distribution between 90 and -10 so Rob is 90% certain that Alice prefers the hotel. Nonetheless Rob has an incentive to defer. The expected utility of booking outright is 40. If Rob proposes the plan and Alice accepts, the expected utility of booking is 45. If Rob proposes the plan and Alice rejects, she receives zero utility. So the expected utility of deferring is $.9 \times 45 + .1 \times 0 = 40.5$, higher than the expected utility of booking outright. However, since Rob is already quite confident about Alice's preferences, the expected utility of deferring is only a little bit higher than the expected utility of booking outright.

The value of deferring looks like an instance of the more general principle that learning is valuable. Hadfield-Menell et al. (2017, p. 222) explicitly draw this analogy: "The reasoning is exactly analogous to the theorem of non-negative expected value of information". As we will see below, the analogy is not perfect since valuing learning and deference can come apart if we allow for misleading signals, but we will first discuss whether AI agents will always value learning.

3 The value of information

Good (1967) shows that if you are an expected utility maximizer, learning is cost-free, you are certain to conditionalize and other assumptions hold, you should always prefer to learn more information before making a decision rather than making the decision without learning.[7]

Here is a quick sketch of Good's theorem. We model your uncertainty by a probability function p on a finite set of states Ω. *Actions* are functions $f : \Omega \to \mathbb{R}$, where $f(\omega)$ is the utility of choosing action f in state ω (Savage, 1972). The *expected utility* of action f relative to probability function p is $\mathbb{E}_p(f) = \sum_{\omega \in \Omega} p(\{\omega\}) f(\omega)$. You learn one element of a partition $\mathcal{E}$ of Ω, where $p(E) > 0$ for all $E \in \mathcal{E}$.

Consider a finite set of actions $\mathcal{S}$. If you choose now, you select one of the actions in $\mathcal{S}$ with maximal expected utility relative to your current credences p, so the expected utility of choosing now is $\max_{f \in \mathcal{S}} \mathbb{E}_p(f)$. We compute the expected value of learning as follows. If you learn $E \in \mathcal{E}$, suppose you are certain to update p by conditionalization to $p(\cdot \mid E)$.[8] Then you choose one of the actions in $\mathcal{S}$ which maximize expected utility relative to your updated credences and receive expected utility $\max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f)$, so the expected value of learning is $\sum_{E \in \mathcal{E}} p(E) \max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f)$. Good (1967) proves that

$$\sum_{E \in \mathcal{E}} p(E) \max_{f \in \mathcal{S}} \mathbb{E}_{p(\cdot|E)}(f) \geq \max_{f \in \mathcal{S}} \mathbb{E}_p(f).$$

which means that if the assumptions of the theorem hold, learning can never make you foreseeably worse off.

Good assumes that learning is cost-free. This means that learning does not affect your set of options and your utility function. The only effect of learning is to change your credences via conditionalization. This might not necessarily be true. For example, Rob's proposal might change Alice's preferences. I set such complications aside but note that they might turn out to be important. For example, we might worry that Rob has an incentive to change Alices preferences so they are easier to satisfy.[9]

4 Rational information aversion

Good's theorem about the non-negative expected value of information makes substantive assumptions. In particular, Good assumes that Rob is an expected utility maximizer and certain to update by conditionalization. There are reasons to be skeptical of both. If these assumptions fail, agents can be required to reject learning. In this case, Rob is

[7] Hosiasson (1931); Blackwell (1951); Howard (1966); Savage (1972); Ramsey (1990) prove similar results. Russell and Norvig (2018, pp. 628-33) discuss the value of information in AI research.

[8] By definition, $p(A \mid E) = \frac{p(A \cap E)}{p(E)}$, assuming $p(E) > 0$.

[9] Russell (2019, p. 139) worries that algorithms which optimize engagement in social media have an incentive to change our preferences so they are easier to satisfy. Even if AI agents allow themselves to be switched off, this problem won't be solved.

not guaranteed to defer to Alice because Rob has no incentive to learn about Alice's preferences.

One assumption is that the agent is an expected utility maximizer.[10] For AI agents following alternative decision theories, learning is not always valuable. One example of such an alternative decision theory is risk-weighted expected utility theory (Buchak, 2010) and other decision theories which relax the independence axiom of expected utility theory (Wakker, 1988). Buchak (2013) argues that such decision theories capture the preferences of many real-life subjects better than expected utility theory. In particular, such decision theories allow agents pay special attention to the worst-case consequences of their actions. It seems reasonable to consider the possibility that we might want to build AI agents which implement such decision theories. However, AI agents implementing such decision theories sometimes prefer to avoid information. Why do agents which care especially about the worst-case scenario avoid information? The reason is that such agents give special weight to the risk of misleading evidence, that is, evidence which suggests that P is true while P is actually false. Buchak (2010) discusses a detailed example.

Another example of decision theories in which learning is not always valuable involve imprecise credences (Kadane et al., 2008; Bradley & Steele, 2016). It seems reasonable to consider such alternative architectures for AI agents. Perhaps we want AI agents to handle situations where we do not have enough information to assign precise probabilities.[11] However, if we go for such alternative architectures, we lose the guarantee that AI agents will always prefer to learn about our preferences.[12]

Good's theorem also requires that the agent is certain that they will update by conditionalization.[13] As Neth (forthcoming) shows, if we allow agents to assign non-zero probability to not conditionalizing, it can sometimes be rational for these agents to reject free information. There are reasons to think that AI agents will assign non-zero probability to violating conditionalization. First, it will be hard to build AI agents which always update by conditionalization because conditionalization is computationally intractable. In particular, Cooper (1990) shows that conditionalization is NP-hard. This means that conditionalization is at least as hard as any problem in the complexity class NP which includes many hard problems like the traveling salesperson problem. Cooper (1990) works in the setting of Bayesian networks which can represent any probability

[10] Bales (2023) critically discusses arguments which claim to show that AI agents will maximize expected utility. I set these arguments aside and focus on reasons why AI agents might follow a different decision theory.

[11] Denoeux et al. (2020) and Caprio et al. (2023) advocate for the use of imprecise probabilities in AI. Ilin (2021) considers a decision theory which allows for ambiguity aversion for applications in autonomous security systems. It is well known that ambiguity aversion leads to information aversion.

[12] As an anonymous referee points out, if designing AI agents following alternative decision theories creates significant risk, perhaps we just shouldn't do it and learn to live with those agents not being sensitive to risk and ambiguity. While this is a reasonable point, many of us are sensitive to risk and ambiguity and might want AI agents to mirror these preferences. If AI agents cannot do so, this is a significant cost.

[13] For example, Skyrms (1990), p. 247 writes that "the proof implicitly assumes not only that the decision maker is a Bayesian but also that he knows that he will act as one. The decision maker believes with probability one that if he performs the experiment he will (i) update by conditionalization and (ii) choose the posterior Bayes act". This means Good's theorem will also fail for agents who are not certain they will maximize expected utility.

distribution over a discrete sample space. In general, if we allow continuous random variables, conditionalization is not even computable (Ackerman et al., 2019). So even if we consider AI agents with lots of computing power, it is not clear whether we can feasibly build them to always conditionalize. At best, AI agents will approximate conditionalization.[14] But approximating conditionalization is not good enough for Good's theorem since any non-zero probability of violating conditionalization leads to some situation where maximizing expected utility requires rejecting information. Similar reasons should make us skeptical that AI agents will maximize expected utility since doing so is also computationally intractable (Bossaerts et al., 2019).

Second, even if we set aside computational limitations, there are general reasons to think AI agents will assign non-zero probability to violating conditionalization. For both human and artificial agents, it seems rational to maintain some uncertainty about how one will update. We are physical systems embedded in the world and many things can go wrong with our updating mechanisms. Sufficiently advanced AI agents will plausibly realize this fact and so assign some probability to failures of conditionalization. So for sufficiently advanced AI agents, Good's theorem does not apply.

There are other well-known limitations of Good's theorem,[15] The upshot is that proponents of CIRL need to pay attention to whether these limitations apply to AI agents. While there is a lot of uncertainty about what AI agents will look like, I have given some reasons to be skeptical whether they will always value learning.

5 Misleading signals

Even if all the assumptions of Good's theorem hold, Rob is not guaranteed to defer to Alice. Hadfield-Menell et al. (2017) assume that Rob has perfect access to Alice's preferences. If this assumption fails and Rob might get Alice's preferences wrong, deferring might not maximize expected utility.

Here is an example to illustrate this point. Like before, Rob is considering whether to book the hotel or defer to Alice. Rob is quite confident that Alice will like the hotel. Rob's uncertainty about how much utility Alice will receive from booking the hotel is given by a uniform probability distribution between 90 and -10.

If Rob defers, it learns whether Alice approves or rejects Rob's proposal. Earlier, we have assumed that Rob is certain that Alice will approve if and only if she prefers the hotel. In other words, we have assumed that Rob has perfect access to Alice's actual preferences, at least in this particular case. Now let us assume that Rob has no such perfect access. There are several reasons for why this might be. For example, Rob might communicate with Alice over a noisy channel with some small probability of garbling Alice's speech, misclassifying 'yes' as 'no' and vice versa. In this case,

[14] Although Dagum and Luby (1993) show that even the problem of approximating conditionalization is NP-hard.

[15] For example, Good's theorem does not apply under either evidential decision theory or causal decision theory if states and actions are correlated (Adams & Rosenkrantz, 1980; Maher, 1990) if evidence does not form a partition (Das, 2023), if probabilities are merely finitely additive (Kadane et al., 1996) and so on.

Rob updates on 'It sounds like Alice said yes' but the probability that Alice actually said yes given that it sounds like Alice said yes is less than one.

Perhaps a noisy channel can be fixed. But there might be deeper obstacles to accessing Alice's preferences. For example, even if Rob can clearly hear what Alice is saying, Rob might nonetheless assign some positive probability to the possibility that Alice is lying or engaged in self-deception. Alice might say she wants the hotel even if she does not really want it. This problem is harder to fix because we *are* sometimes lying about our preferences, even to ourselves. Furthermore, humans might send misleading signals for strategic reasons. So it seems very plausible that advanced AI agents will assign some probability to humans sending misleading signals. But as we will see in a moment, this means that Rob is not guaranteed to defer.

Is there any way to block this move by hard-wiring AI agents to ignore the possibility of misleading signals? Depending on how the AI agent is trained, this might not be feasible. More importantly, it seems like an accurate model of human psychology is important for AI agents to perform well in many real-world tasks such as playing *Diplomacy*.[16] An AI agent which assigns probability zero to humans sending misleading signals about their preferences will be seriously handicapped in such tasks and so seems unlikely to be deployed in critical real-world applications.[17]

We can modify the Off-Switch Game to allow for misleading signals as follows. Like before, if Rob defers, Alice will either approve or reject Rob's proposal. But now with some small probability $\epsilon > 0$ Alice's signal is misleading and does not reflect her true preferences.[18] In particular, given that Alice prefers the hotel, with probability ϵ she rejects Rob's proposal. And given that Alice does not prefer the hotel, with probability ϵ she approves Rob's proposal. Rob's new decision problem is depicted in Fig. 3. Like before, the expected utility of booking outright is 40. If Alice really prefers the hotel, the expected utility of booking after deferring is 45. But if Alice sent an incorrect signal, the expected utility of booking is -5. If ϵ is bigger than around 1.2%, Rob maximizes expected utility by not deferring but instead booking without asking.[19]

Note that as I've described the case, Rob satisfies all the assumptions of Good's theorem. This means that Rob will always value learning more information before making a decision. How is this compatible with Rob not deferring? The answer is that deferring is not the same as learning more information before making a decision.

[16] Thanks to Aydin Mohseni for suggesting this example.

[17] Failures of deference can also arise from 'model misspecification' (Milli et al., 2017; Carey, 2018) so trying to ensure deference by intentionally giving the AI agent an inaccurate model of human psychology is not a promising idea.

[18] We can think of Alice's signal as a partition of our state space Ω. As Ye (forthcoming) explains, following Blackwell (1951), we can also think of signals as inducing a probability distribution on Ω. From this perspective Hadfield-Menell et al. (2017) assume the signal perfectly reveals Alice's preferences while we study a 'garbled' signal with added noise.

[19] The expected utility of deferring is $p(\text{approve})\big(p(\text{prefer} \mid \text{approve})45 - 5p(\text{dispref} \mid \text{approve})\big)$. We can calculate $p(\text{approve}) = p(\text{dispref})\epsilon + p(\text{prefer})(1-\epsilon) = 0.1\epsilon + 0.9(1-\epsilon)$. By Bayes' theorem, $p(\text{prefer} \mid \text{approve}) = p(\text{prefer})\frac{p(\text{approve}|\text{prefer})}{p(\text{approve})} = 0.9\frac{1-\epsilon}{0.1\epsilon+0.9(1-\epsilon)}$ and $p(\text{dispref} \mid \text{approve}) = 1 - p(\text{prefer} \mid \text{approve})$. We can calculate $p(\text{prefer} \mid \text{reject})$ and $p(\text{dispref} \mid \text{reject})$ analogously. See Jupyter notebook available at https://github.com/nethsv/off-switching for details.

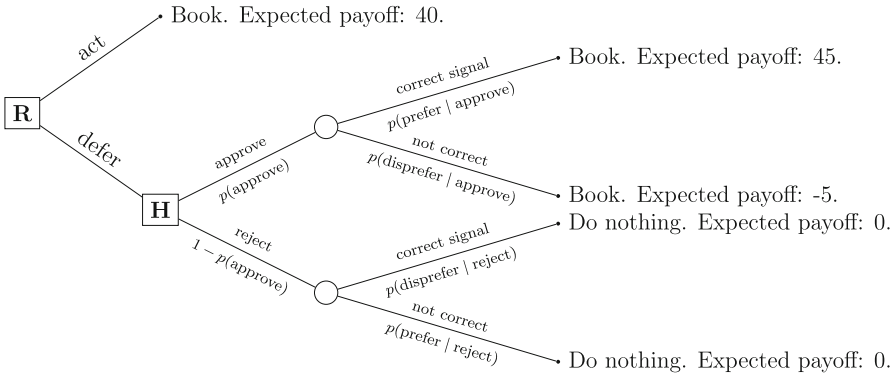

Fig. 3 Rob's decision problem with uncertain access to preferences

When choosing to defer, Rob does whatever action Alice proposes. So while deferring gives Rob new information, it also changes Rob's choice set. For example, Rob is not able to learn that Alice rejects the proposal but then still implement it. If Rob had the option to learn Alice's signal and then still choose among the original option set {book, do nothing}, this option would always maximize expected utility. So the link between preferring to learn new information and deference is weaker than Hadfield-Menell et al. (2017) suggest. For example, if $\epsilon = 0.02$, the expected value of deference is slightly less than 40 and the expected value of learning is 40 since Rob will book no matter which signal Alice sends.[20] This shows that preferring to defer and preferring to learn more information can come apart. Instead of booking outright, Rob might also listen to Alice's signal but then proceed ignore her. This doesn't make the situation any better.[21]

You might complain that this example is unrealistic. Perhaps a probability of more than 1% of Alice sending an incorrect signal is too pessimistic. But we can construct a similar example for any non-zero probability of misleading signals if we make Rob even more confident that booking is right. More broadly, the example illustrates a general lesson. If Rob is not perfectly certain of learning Alice's actual preferences, there is no guarantee that Rob will defer. It might still be true that Rob defers to Alice most of the time. But for provably beneficial AI, this is not enough. If we allow the possibility of misleading signals, off-switching is not guaranteed.

[20] See Jupyter notebook available at https://github.com/nethsv/off-switching.

[21] If deferring means that Rob implements whatever Alice proposes, does this put pressure on our earlier arguments about information aversion? It depends on how we construe the Off-Switch Game. If Alice's signaled preference is implemented automatically, Rob might refuse to defer because of the possibility of misleading signals. If Alice's signaled preference isn't implemented automatically and Rob maximizes expected utility after learning, Rob might refuse to defer because of negative information value. In this case misleading signals don't lead Rob to refuse to defer, but 'deferring' might mean learning Alice's signal and then ignoring it. In the original Off-Switch Game, it doesn't make a difference whether or not Alice's preference is implemented automatically, these cases only come apart when we consider information aversion and misleading signals. Thanks to an anonymous referee for pressing me on this point.

You might argue that if Rob is uncertain about whether it will learn our actual preferences, then it should not always defer to us.[22] There is clearly a sense in which Rob is acting rationally by not deferring. Rob assigns high prior probability to booking being right so from Rob's point of view, when Alice rejects, it is relatively likely that Alice sent a misleading signal. Together with the potential downsides of failing to book, this means that Rob maximizes expected utility by not deferring. This reasoning looks acceptable from our 'outside' point of view if Rob has a reasonable prior. But if Rob has a strange prior, this reasoning can look really bad. This is a problem because one of the core motivations for the Off-Switch Game is that it is very hard to specify a reward function which correctly captures our preferences. But, the idea goes, we don't have to worry about specifying the correct reward function because Rob will always defer to us. Presumably it is equally hard to specify the correct prior over our preferences. But in the presence of misleading signals, we *do* have to worry about Rob's prior. Depending on its prior, Rob might or might not let itself be switched off. Thus, the Off-Switch Game does not successfully solve the problem it was designed to solve: making sure AI agents defer to us without having to worry about the details of their reward function and prior. Given that the assumptions of Good's theorem hold, AI agents still have an incentive to 'listen to us' but, depending on their prior, they might decide to ignore our signals. This seems like cold comfort.

Now suppose the stakes are higher. Rob is contemplating a permanent change to our environment, perhaps a plan to stop climate change which, as side effect, permanently turns the sky orange (Russell, 2019, p. 202). Rob is quite confident that this is the right option and has some uncertainty about whether it will learn our actual preferences, so it goes ahead and implements this plan without asking. This seems like a situation where we really want Rob to defer to us. If Rob has an incentive to not defer, this is a problem.

6 Practical significance

You might respond as follows. Grant that there is a possibility AI agents will not defer to us. But how likely is this possibility? It might be that AI agents defers to us in almost all cases. Should we still be worried about the tiny minority of cases where they fail to defer?[23]

There are two responses. First, even one instance of non-deference might have catastrophic consequences. AI agents might have powerful capacities to change our world in irreversible ways. This means that once the change is rolled out, it might be impossible to roll it back. One of the most worrying examples for an irreversible change is human extinction. But there are less drastic examples as well. AI agents might use up some non-renewable resource, permanently change our preferences or permanently turn the sky orange.

[22] Thanks to an anonymous referee for raising this objection. Milli et al. (2017) also argue that when humans are behaving irrationally, AI agents should not always defer.

[23] Thanks to an anonymous referee for raising this concern.

Second and more importantly, we don't have good reasons to think that the probability of non-deference is extremely low. Phenomena like sending misleading signals about one's preferences seem widespread. It seems implausible to ignore such phenomena when considering whether AI agents will defer to us in real life. So it seems that there is not just an in-principle possibility but a substantial probability that AI agents will fail to defer. It is difficult to see why we should be confident that in practice, non-deference is extremely unlikely. Perhaps proponents of CIRL could provide an argument for why this is true. This argument would need to show that, among other things, AI agents are very likely to maximize expected utility, be certain of updating by conditionalization and have perfect access to our preferences. Why is that? This is an important question which proponents of CIRL should address.

7 A dilemma for provably beneficial AI

I have argued that the result of Hadfield-Menell et al. (2017) relies on substantive assumptions. Even if we make AI agents uncertain of our preferences, it is not guaranteed that they will always defer to us. This highlights a more general dilemma for provably beneficial AI. To prove that AI agents always defer to us (or are beneficial in some other sense), you have to make some decision-theoretic and epistemological assumptions. Either you make strong or weak assumptions.

Strong assumptions, such as expected utility maximization, certain conditionalization and perfect access to our preferences, will allow you to prove interesting guarantees. But such assumptions might not apply to all AI agents. As we have seen, there are reasons to think that AI agents in the real world might not be certain of updating by conditionalization and have perfect access to our preferences. So even if you can prove that given such assumptions, AI agents will be beneficial, this isn't much comfort if the assumptions might not be satisfied by many AI agents. Weak assumptions apply to a wider range of possible AI agents but don't allow you to prove much. As we have seen, if AI agents assign some positive probability to failures of conditionalization or to humans sending misleading signals about their preferences, they are not guaranteed to defer to us.

Perhaps there is a way to successfully navigate this dilemma. We might find decision-theoretic assumptions weak enough to cover (almost) all plausible varieties of AI agents and strong enough to prove interesting guarantees—assumptions which are 'just right'. But since we know so little about what AI agents might look like, it is not clear whether this will work out.

8 Conclusion

Hadfield-Menell et al. (2017) propose a model for making sure that AI agents defer to us by making them uncertain about our preferences. I have argued that their result relies on strong decision-theoretic and epistemological assumptions: AI agent maximize expected utility, are certain of updating by conditionalization and have perfect access

to our preferences. These assumptions limit the scope of the model since they might not be satisfied by AI agents in the real world.

Everything I have said here is compatible with the broad idea that we shouldn't tell AI agents to maximize a particular reward function but rather 'teach them as we go along'. But we need more solid foundations for this idea. The problem of making sure AI agents defer to us is not yet solved.

Acknowledgement Thanks to Mel Andrews, Mathias Böhm, Lara Buchak, Shamik Dasgupta, Daniel Filan, John MacFarlane, Aydin Mohseni, Emily Perry, David Thorstad, two anonymous referees and an editor of *Philosophical Studies*, and audiences at the 7th Annual CHAI Workshop in Pacific Grove, CA for very helpful discussions.

Declarations

Conflict of interest Author declares that they have no Conflict of interest to declare.

References

Ackerman, N. L., Freer, C. E., & Roy, Daniel M. (2019). On the computability of conditional probability. *Journal of the ACM, 66*(3), 1–40. https://doi.org/10.1145/3321699

Adams, E. W., & Rosenkrantz, R. D. (1980). Applying the Jeffrey decision model to rational betting and information acquisition. *Theory and Decision, 12*(1), 1–20. https://doi.org/10.1007/BF00154655

Andrews, M., Smart, A., & Birhane, A. (2024). The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns, 5*(9), 1–14. https://doi.org/10.1016/j.patter.2024.101027

Bales, A. (2023). Will AI avoid exploitation? artificial general intelligence and expected utility theory. In: *Philosophical Studies*. https://doi.org/10.1007/s11098-023-02023-4.

Blackwell, D. (1951). Comparison of experiments. In: *Proceedings of the second berkeley symposium on mathematical statistics and probability*, Vol. 2. Berkeley and Los Angeles: University of California Press, pp. 93–103.

Bossaerts, P., Yadav, N., & Murawski, C. (2019). Uncertainty and computational complexity. *Philos Trans R Soc B, 374*(1766), 1–12. https://doi.org/10.1098/rstb.2018.0138

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

Bradley, S., & Steele, K. (2016). Can free evidence be bad? Value of information for the imprecise probabilist. *Philosophy of Science, 83*(1), 1–28. https://doi.org/10.1086/684184

Buchak, L. (2010). Instrumental rationality, epistemic rationality, and evidence-gathering. *Philosophical Perspectives, 24*(1), 85–120. https://doi.org/10.1111/j.1520-8583.2010.00186.x

Buchak, L. (2013). *Risk and rationality*. Oxford University Press.

Caprio, M., Dutta, S., Jang, K. J., Lin, V., Ivanov, R., Sokolsky, O., & Lee, I. (2023). Credal Bayesian deep learning. In: arXiv preprint arXiv:2302.09656. https://doi.org/10.48550/arXiv.2302.09656.

Carey, R. (2018). Incorrigibility in the CIRL Framework. In: *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society*, pp. 30–35. https://doi.org/10.1145/3278721.3278750.

Cooper, Gregory F. (1990). The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence, 42*(2–3), 393–405. https://doi.org/10.1016/0004-3702(90)90060-D

Dagum, P., & Luby, M. (1993). Approximating probabilistic inference in bayesian belief networks is NP-hard. *Artificial Intelligence, 60*(1), 141–153. https://doi.org/10.1016/0004-3702(93)90036-B

Das, N. (2023). The value of biased information. *British Journal for the Philosophy of Science, 74*(1), 25–55. https://doi.org/10.1093/bjps/axaa003

Denoeux, T., Dubois, D., & Prade, H. (2020). Representations of uncertainty in AI: beyond probability and possibility. In: *A guided tour of artificial intelligence research*, Vol. I. Springer, pp. 119–150. https://hal.science/hal-02921351/file/volume-1-chapitre-4-Springer.pdf.

Gallow, J. D. (2024). Instrumental divergence. *Philosophical studies*. https://doi.org/10.1007/s11098-024-02129-3.

Good, I. J. (1967). On the principle of total evidence. *British Journal for the Philosophy of Science, 17*(4), 319–321. https://doi.org/10.1093/bjps/17.4.319

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). Cooperative inverse reinforcement learning. *Advances in Neural Information Processing Systems, 29* .

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The off-switch game. In: *Proceedings of the 26th international joint conference on artificial intelligence, IJCAI-17*, pp. 220–227. https://doi.org/10.24963/ijcai.2017/32.

Hosiasson, J. (1931). Why do we prefer probabilities relative to many data? *Mind, 40*(157), 23–36. https://doi.org/10.1093/mind/xl.157.23

Howard, R. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics, 2*(1), 22–26. https://doi.org/10.1109/TSSC.1966.300074

Ilin, R. (2021). Detection of rare events with uncertain outcomes. *International Journal of Approximate Reasoning, 131*, 252–267. https://doi.org/10.1016/j.ijar.2020.12.022

Kadane, J. B., Schervish, M., & Seidenfeld, T. (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association, 91*(435), 1228–1235. https://doi.org/10.1080/01621459.1996.10476992

Kadane, J. B., Schervish, M., & Seidenfeld, T. (2008). Is ignorance bliss? *Journal of Philosophy, 105*(1), 5–36. https://doi.org/10.5840/jphil200810518

Maher, P. (1990). Symptomatic acts and the value of evidence in causal decision theory. *Philosophy of Science, 57*(3), 479–498. https://doi.org/10.1086/289569

Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). Should robots be obedient? In: Sierra, C. (Ed.) *IJCAI'17: Proceedings of the 26th international joint conference on artificial intelligence*, pp. 4754–4760. https://doi.org/10.24963/ijcai.2017/662.

Neth, S. (forthcoming). Rational aversion to information. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/727772.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Bloomsbury.

Ramsey, F. (1990). Weight or the value of knowledge. *British Journal for the Philosophy of Science, 41*(1), 1–4. https://doi.org/10.1093/bjps/41.1.1

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

Russell, S., & Norvig, P. (2018). *Artificial intelligence: A modern approach*. Third Edition. Prentice Hall.

Savage, Leonard J. (1972). *The foundations of statistics*. Second Revised Edition: Wiley Publications in Statistics.

Skyrms, B. (1990). The value of knowledge. In: Savage, C. W. (Ed.) *Scientific theories*, Vol. 14. Minnesota Studies in the Philosophy of Science. University of Minnesota Press, pp. 245–266.

Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). Corrigibility. In: *Workshops at the twenty-ninth AAAI conference on artificial intelligence*. https://cdn.aaai.org/ocs/ws/ws0067/10124-45900-1-PB.pdf.

Thorstad, D. (2024). Against the singularity hypothesis. *Philosophical Studies*. https://doi.org/10.1007/s11098-024-02143-5.

Wakker, P. (1988). Nonexpected utility as aversion of information. *Journal of Behavioral Decision Making, 1*(3), 169–175. https://doi.org/10.1002/bdm.3960010305

Ye, R. (forthcoming). The value of evidence in decision-making. *Journal of Philosophy*.

Zhuang, S., & Hadfield-Menell, D. (2020). Consequences of misaligned AI. *Advances in Neural Information Processing Systems, 33*, 15763–15773.