

Off Switch Game- Rapporteur Report

Praneel Seth

October 2026

1 Paper Summary

The Off-Switch Game was designed to explore when a rational agent would have an incentive to preserve a human's ability to shut it down (keep an off switch) [1]. The motivation comes from the idea that self-preservation can emerge as an instrumental goal: even if an agent is not explicitly rewarded for staying alive, being shut down prevents it from continuing to optimize its objective. Earlier work on corrigibility tried to address this by making an agent indifferent between shutdown and continuing to optimize its original utility function [2]. Instead of forcing indifference, this paper asks whether uncertainty about the true objective can make preserving the off switch positively valuable to the agent.

The paper models the problem as a game between a robot R and a human H . The robot can take an action a immediately, shut itself down with action s , or choose $w(a)$, which means announcing its intended action and allowing the human to either permit the action or shut the robot down. The utility of shutdown is defined as zero. The true human utility of taking a is U_a . A key assumption is that the human knows U_a , while the robot does not. Instead, the robot has a belief distribution $P(U_a)$ over the possible values of the action.

Because the robot does not know U_a , it must evaluate its choices using expected utility. Directly taking a has value $E[U_a]$, while shutting down has value zero. The important question is therefore the value of $w(a)$. If the human's shutdown decision were random and independent of U_a , waiting for the human could not improve on simply choosing the better of a and shutdown. It would only produce some weighted average between U_a and zero. The idea that makes this paper work is that an actual human decision can instead contain information about U_a .

For a perfectly rational human, the human allows the action whenever $U_a \geq 0$ and shuts the robot down whenever $U_a < 0$. This is a strong assumption that the U_a coded inside the model as its goal is always what the human wants to optimize as well. In this case, human oversight filters the robot's outcomes: positive-utility actions are allowed through, while negative-utility actions are replaced by shutdown. For example, the robot could assign probability 0.8 to $U_a = 10$ and probability 0.2 to $U_a = -20$. Taking the action directly therefore has expected value 4, while allowing the human to intervene gives expected value 8, because the human removes the negative outcome.

The paper formalizes the value of preserving human oversight using

$$\Delta = E[\pi^H(U_a)U_a] - \max\{E[U_a], 0\},$$

where $\pi^H(U_a)$ is the probability that the human allows the action. Under the perfectly rational human policy, Δ is always nonnegative. If the robot assigns positive probability to both positive and negative values of U_a , then Δ is strictly positive: the human has information that can improve the robot's decision. You can think of this as $w(a)$ filtering out bad outcomes while retaining good ones. The paper then relaxes the assumption that humans are perfectly rational. Human decisions are modeled as increasingly noisy, while the robot's uncertainty over U_a is represented through the variance of its belief. This creates a balance between the value of human information and the cost of allowing an unreliable human to override the robot's own estimate. As a result, human oversight is most valuable when the robot is uncertain and the human is informative, but it may cease to be optimal when the robot is confident, and the human is sufficiently unreliable. The overall conclusion is that uncertainty about the objective can create a rational incentive to preserve human oversight, but only when human behavior

provides useful information. Essentially, the moment the robot begins to lose confidence in the human's ability to provide oversight that matches its U_a , it is less likely to preserve the off switch.

2 My Thoughts

What I found most interesting about the paper was the way it uses uncertainty to produce corrigible behavior. Rather than directly modifying the utility function so that the agent is indifferent to shutdown, the paper shows that preserving human control can arise naturally from expected-utility maximization. If the robot recognizes that it may be wrong about the true utility of an action, then the human's decision becomes a source of information. The off switch can therefore actually have positive expected value from the agent's own perspective. I also liked how simple the central argument becomes once the setup is clear. The result is essentially an application of the value of information. If the human knows something relevant to the decision that the robot does not, then allowing the human to intervene cannot hurt under the assumption of perfect human rationality.

At the same time, I think the simplicity of the model is also its main limitation. The result depends on strong assumptions about the relationship between the human, the robot, and the objective. The robot is already assumed to be cooperatively trying to maximize human utility, and the uncertainty is about what that utility is, not about whether the robot is actually aligned with the human objective in the first place. In that sense, the paper assumes much of the outer-alignment problem is already solved. It does not address the harder case where the objective encoded into the model is itself wrong, incomplete, or learned in a way that causes the system to pursue something other than what humans intended. We don't have uncertainty built in for the model of U_a that the agent has. Therefore, a human that takes shutdown while trying to maximize actual human utility might look irrational from the perspective of an agent that thinks its utility function is correct.

I therefore see the paper less as a complete solution to corrigibility/keeping shut down available and more as an important conceptual result. It shows that uncertainty can change an agent's incentives in a useful direction. Instead of always trying to remove uncertainty from an AI system, there may be safety value in maintaining calibrated uncertainty about whether the system has correctly understood the human objective. However, turning this theoretical result into a modern learned system would require solving much harder problems about how objectives are represented, how uncertainty over those objectives is represented, and how reliable human feedback actually is.

3 Class Discussion

The class discussion focused heavily on the assumptions behind the formal model and on how much of the result follows from those assumptions.

One early question concerned the paper's use of a Gaussian distribution for the robot's uncertainty over U_a . The class noted that the conceptual result does not seem to depend strongly on the distribution being Gaussian. A much simpler distribution with only two possible values—one where the action is good and one where it is bad could also capture the representation we are looking for. The Gaussian assumption becomes more useful later because the paper uses the mean and variance analytically, but it is not necessary to understand the basic value-of-information argument.

Another important point was that under the perfectly rational human assumption, $w(a)$ is almost trivially optimal if consulting the human has no cost. The class asked why the agent would not simply ask the human before every action. Under the model's assumptions, if the human always knows the correct utility, always makes the optimal decision, and there is no cost to asking, more information cannot hurt the agent. This led to the observation that a realistic system would almost certainly have some cost associated with human oversight, such as time, attention, or the impracticality of requesting approval for every action. The more interesting part of the paper therefore begins when the assumption of perfect human rationality and availability is relaxed. Where's the term for the cost of waiting for a human to reply?

The discussion also emphasized that the core theorem is closely related to the general value-of-information principle: receiving useful information before making a decision should not reduce expected value. This helped clarify why the main mathematical result can feel obvious once the information structure of the game is understood.

Finally, an analogy involving Rapunzel led to a useful distinction between uncertainty and actual preference misalignment. The initial analogy treated Rapunzel as the agent and her mother as the better-informed human, but the class pointed out that their preferences are genuinely misaligned rather than merely uncertain. A parent-child example in which both share the same underlying goal but the parent has better information is closer to the assumptions of the paper. This discussion highlighted an important limitation of the model: the off-switch result depends on the robot and human ultimately sharing the same objective, with the disagreement arising from uncertainty or unequal information rather than fundamentally different goals. Even if the agent has some distribution representing the utility function, either a misaligned human will be represented as wildly irrational, or a model training based on the human's decisions may become misaligned itself under this paradigm.

Overall, the discussion suggested that uncertainty can be a useful mathematical property to make agents more responsive to humans, but this paper has narrow applications today. Under the assumptions of shared objectives, useful human information, and low-cost oversight, the incentive to preserve the off switch follows naturally. The more difficult alignment questions appear when those assumptions are weakened.

References

- [1] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, "The off-switch game," in Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17), Melbourne, Australia, 2017, pp. 220–227.
- [2] E. Yudkowsky, B. Fallenstein, N. Soares, and S. Armstrong, "Corrigibility," Machine Intelligence Research Institute, Berkeley, CA, Tech. Rep., 2014.