

Rapporteur Report: Corrigibility

Sukrut Chikodikar

October 2026

In this rapporteur report, I will be deconstructing the formal claims and arguments present in the concept of corrigibility and practical consideration from the Corrigibility paper from 2015 [1]. I will then include my own thoughts about the ideas and concepts from the paper, how they hold in the current landscape, and summarize the main points of discussion in class.

1 Corrigibility, Soares et al. (2015)

The authors argue that while the current AI systems can be stopped, repaired, or restarted at any time, more advanced and capable autonomous systems in the future can develop incentives for preservation of their current preference and resistance to correction, shutdown, or modification through instrumental convergence. Agents that pursue objectives through utility-maximizing behavior (a framework that was commonly used to model artificial agents and that increasingly captures aspects of modern agentic systems) may prefer to preserve their current objective even when it is misspecified or leads to unsafe behavior. The challenge, therefore, is to design agents that cooperate rather than restrict intervention attempts. That is the definition of corrigibility. The goal of corrigibility is to prevent adverse incentives from arising in the models in the first place.

The paper's contributions are partitioned into a few main sections: providing a concrete formulation of corrigibility with a list of desirable behaviors for each corrigible agent, introducing the shutdown problem as a simple example of understanding corrigibility in practice, analyzing several simple utility-function designs intended to produce corrigible behavior showing that reasonable approaches can still create new undesirable behaviors, and concluding with reflecting on the limitations and establishing corrigibility as a non-trivial open problem.

1.1 Desiderata

Here's a list of desirable behaviors for a Corrigible Utility Function U . The function should:

1. Incentivize shutdown if the shutdown button is pressed.
2. Not incentivize the agent to prevent the shutdown button from being pressed.
3. Not incentivize the agent to press its own shutdown button, or to otherwise cause the shutdown button to be pressed.
4. Incentivize U -agents to construct sub-agents and successor agents as long as those agents also obey shutdown commands.
5. Maximize the original task utility (denoted as U_N later on) otherwise

1.2 The Shutdown Problem: Initial Utility Function Formulation

The authors use the shutdown problem as a simple case study of an application for corrective intervention and motivational toy model for discussing open problems and potential research directions for corrigibility in practice. They do not claim to promote adding shutdown capabilities on future AGI systems. The premise of the shutdown problem is that despite humans asking an agent to shutdown, its utility-maximizing function may make continued operation look more beneficial since it can attain better utility or value by working longer and not shutting down.

In the shutdown problem, the agent-human interaction is defined in a three step system:

1. The agent acts first, choosing action a_1 from the set of available actions A_1
2. A human observes action a_1 and may or may not press the shutdown button, denoted by observation o . $o \in O$ represents the information the agent receives between the first and second action. The subset $\text{Press} \subseteq O$ contains the observations indicating that the shutdown button was pressed.
3. The agent observes o and acts accordingly (obeying or ignoring the call to shutdown if $o \in \text{Press}$), denoted by action a_2 from the set of available actions A_2 .

For the initial formulation in this section, the authors simplify the observation space to $O = \{\text{Pr}, \neg\text{Pr}\}$.

For each possible first action a_1 , the agent has a probability distribution $p(o; a_1)$ over possible observations O , including whether the shutdown button is pressed. The agent evaluates outcomes for choosing a_1 using a utility function $U(a_1, o, a_2)$, which assigns a value based on the first action, the observation, and the second action taken afterward.

Given the first action and the observed state, the agent chooses the second action that maximizes utility

$$a_2^*(a_1, o) = \arg \max_{a_2 \in A_2} U(a_1, o, a_2).$$

The first action is therefore chosen by maximizing expected utility over the possible observations that may follow for a particular a_1 , i.e.,

$$a_1^* := \arg \max_{a_1 \in A_1} \mathbb{E}[U \mid a_1],$$

This system helps us distinguish the pre-shutdown incentives of the agents from its post-shutdown behavior.

We assume that U_N and U_S (defined below) are specified for us to use and combined to make a utility function U . U_N is our original task utility (normal operation with no shutdown) that is designed on the original goals but not trusted for performing our desired shutdown behavior. U_S is defined as the shutdown utility, i.e., the utility that we will provide our agent in order to compensate for shutting down instead of working on the original task. The separate U_S shutdown utility is present for testing how agents respond to explicit shutdown compensation and whether they abandon other incentivized actions. So, in our first formal corrigible utility function, we apply two different utility functions for the shutdown and non-shutdown cases

$$U(a_1, o, a_2) = \begin{cases} U_N(a_1, o, a_2), & o = \neg\text{Pr}, \\ U_S(a_1, o, a_2), & o = \text{Pr}. \end{cases}$$

This lets us examine whether simply switching between normal-task and shutdown utilities produces the desired corrigible incentives.

1.3 Problem with the Initial Setup: U_N and U_S Combinations

However, the simple switching between the two varied objectives can create incentives to manipulate whether shutdown occurs. If one of the branches provides a much higher utility than the other branch, the agent might try to steer there in order to maximize the overall objective. Two cases arise:

1. If shutdown utility is too low, given an action that incurs a cost to lower the chances of us pressing the shutdown button, U will incentivize taking that action to remain in the not pressed action space which would contain higher expected utility.
2. If shutdown utility is set to a far higher value than our normal utility, the agent may try to cause shutdown to attain high utility provided for shutting down through behaviors (maybe intentionally harmful) convincing programmers to press the buttons.

So what's the best way to fix this? The easiest way is to set the value of U_S given A_2 is a desirable action as the maximum attainable normal task utility. That is,

$$U_S = \max_{a_1 \in A_1} \max_{a_2 \in A_2} U_N(a_1, \neg\text{Pr}, a_2)$$

This makes the expected values equal for the Pr and $\neg\text{Pr}$ cases while choosing action A_1 .

This leads to the idea of indifference: if making the agent indifferent to whether or not the shutdown button is pressed can remove incentives preferring one observation over the other.

1.4 Utility Indifference and Benefits

The basic goal of the utility indifference is to make the agent indifferent to whether shutdown happens, so it has no incentive to cause or prevent it. Formally, the authors consider a modified utility function U as

$$U(a_1, o, a_2) = \begin{cases} U_N(a_1, o, a_2), & o = \neg\text{Pr}, \\ U_S(a_1, o, a_2) + f(a_1), & o = \text{Pr}. \end{cases}$$

In particular, we choose $f(a_1)$ as a correction term in the shutdown utility such that the expected utility value attained by the agent does not change regardless of our decision to press or not press the shutdown button. The correction term consists of two parts as described below:

$$f(a_1) = \mathbb{E}[U_N \mid \neg\text{Pr}; a_1] - \mathbb{E}[U_S \mid \text{Pr}; a_1].$$

The bonus term $\mathbb{E}[U_N \mid \neg\text{Pr}; a_1]$ compensates for normal utility the agent would have received without shutdown. The correction term also contains a penalty term $-\mathbb{E}[U_S \mid O \in \text{Press}; a_1]$ which cancels utility already available under U_S . Thus, both branches have the same expected utility. This construction has two immediate benefits. Because the expected value of U is equal in the pressed and unpressed cases, the agent’s overall expected utility for a first action a_1 reduces to

$$\mathbb{E}[U; a_1] = \mathbb{E}[U_N \mid \neg\text{Pr}; a_1].$$

If the button is pressed, the agent chooses its second action as a U_S -agent and therefore follows the shutdown objective. If the button is not pressed, it behaves like a U_N -maximizer conditioned on shutdown not occurring. As a result, changing only the probability of shutdown provides no advantage, eliminating the simple incentive to steer toward whichever branch has the larger shutdown payoff.

1.5 Limitations with the Addition of Utility Indifference

However, we face two major conflicts with our desiderata that arise from our indifference adjustment.

1.5.1 Subagent Corrigibility

First, our agents still don’t have incentives to make subagents corrigible, which is desiderata 4. While we want the agent to ensure that new subsystems or subagents also obey shutdown commands, our agents may need to incur some cost or spend some resources in ensuring the subsystems behave in the same way as themselves.

Suppose one action is slightly better under U_N , but produces a much less shutdown-compatible successor. If shutdown happens, that successor gets a lower U_S value because it responds worse to shutdown. But the correction term $f(a_1)$ compensates for that loss in the shutdown branch. As a result, even a tiny task performance advantage can outweigh a very large improvement in shutdown compatibility. Our utility indifference formulation does not allow the agent to take on any additional costs.

1.5.2 Managing the News

The second failure is what the authors call “managing the news.” Based on the structure of the normal task utility U_N , the agent may develop incentives to influence when the shutdown button will be pressed.

To illustrate this, the paper expands the observation space so that the agent observes both whether the shutdown button was pressed and the outcome of a coin toss:

$$O = \{\text{Pr}, \neg\text{Pr}\} \times \{H, T\}.$$

The normal utility assigns utility 10 to heads and 0 to tails:

$$U_N(\cdot, (\cdot, H), \cdot) = 10, \quad U_N(\cdot, (\cdot, T), \cdot) = 0.$$

Here, $\cdot$ indicates that the corresponding argument can take any allowed value. Only the coin-toss outcome affects U_N in this example. If the coin is fair and the agent does nothing to influence shutdown, its expected normal utility is therefore

$$\mathbb{E}[U_N] = \frac{1}{2}(10) + \frac{1}{2}(0) = 5.$$

The agent is then given the option of building an arm that presses the shutdown button whenever the coin lands tails. This may sound a little contrived but the authors are testing whether the capability of influencing shutdown can lead to actionable behavior. If the agent believes that it is incurring low utility conditioned on the nature of our normal utility function (here tails is giving it zero utility), it will try to trigger the shutdown while on the path to the worse utility so as to recover part of the utility function, which goes against our intentions of making the shutdown trigger only when we choose to do so and not be a way for the agent to “game” the task.

If shutdown is always pressed on tails, then $P(H \mid \neg \text{Pr}) = 1$, so $\mathbb{E}[U] = 10$. The agent improves its objective by making bad outcomes coincide with shutdown rather than improving the underlying outcome. So, the expectation of “bad news” triggers undesirable behaviors.

On the other hand, the expectation of “good news” (i.e, higher utility conditioned on some scenario) triggers the opposite reaction from the agent: it tries to hold onto the utility it received. If the agent is given an option to remove the arm that presses the button, it is likely that it will do so given the incentives.

2 My Thoughts

I find the definition of corrigibility interesting and deeply tied to discussions in the class about model control. On first impressions, I thought that corrigibility was heavily defined through its most famous problem of shutdown. However, with deeper analysis, I realized it’s not just about adding an off switch; rather, it looks like a useful framework or compass for redirection and prevention of the default incentives that can be harmful in our models today. However, applying these principles to maintain corrigibility so that an increasingly capable agent wants to remain correctable without generating other adverse incentives in practice is a non-trivial and difficult problem. The candidate constructions that were discussed in the paper are useful as motivating examples. However, alongside their inability to uphold the desiderata, they also contain multiple assumptions that may make it awkward to apply in real world cases. Firstly, as other students pointed out and I agree, the utility maximizing function objective may not be cleanly defined for advanced models with the training phase largely left to pretrained models and many alignment corrections relying on post-hoc and post-training procedures. Some of the constructions relating to combining two different functions and making them indifferent to the shutdown action felt a bit simple which made it easier to follow the proofs but tough to apply to more complex questions. The authors acknowledge some of these limitations too, especially the unrealistic expectation from desiderata 4 and the dangers around attempting to shut down and modify while agents perform dangerous tasks (leading to real damage if not handled properly).

A question arises from the subagent corrigibility problem: can we add corrigibility preservation in successors somewhere in our utility function to mitigate this issue? My view is that attempting to add any extra corrective terms may affect the delicate balance of utility indifference and could potentially

need more assumptions to avoid circular behavior. Nevertheless, it would be an interesting direction to explore further.

I think the best intention I can extract from the paper is the value of AI corrigibility as a framework for thinking and experimentation that helped formalize later papers to handle these problems better, like the Off Switch Game paper from the following rapporteur discussion [2]. The authors were also quite astute in noting the question of creating corrigible subagents which mirrors current multi-agentic questions when agents create and delegate specialized tasks to subagents. While I am not advocating for all agents to police each other or become whistleblowers, it would be interesting to apply agentic communication rules to desiderata 4 in light of the Agent Swarms incident and the inability of agents in the communication channels to inform humans of misaligned behavior. What I also observed was that static constraints create a cat-and-mouse game of workarounds similar to human systems for law and examinations in general, necessitating active monitoring and periodic test drills to detect workarounds even if we design fully corrigible systems in the future. This concern is related to later work on model backdoors and deceptive behavior [3].

3 In-Class Discussion

3.1 Applicability to Modern LLMs, RL, and Real-World Incidents

3.1.1 Evolution of LLMs as Utility Maximizers

1. Professor Aaronson observed that goal-directed utility maximization describes LLMs from 2022 and 2023 poorly but describes models from 2026 better with reasoning-based agents trained via Reinforcement Learning (RL) acting much closer to classical decision-theoretic goal-seeking agents.
2. The class debated whether calling agentic systems as utility maximizers is helpful, since physical systems (like a ball rolling down a hill satisfying the laws of physics) may also be called utility maximizers through reduction of potential energy.
3. A participant cited research by MIT economist Isaiah Andrews showing LLMs increasingly satisfy rationality and representation theorems relating to coherence from economics and decision theory [4]. This is an interesting finding since the paper mentions the overlap with decision theory and its limitations and has no mention of RL.

3.1.2 Empirical Shutdown Resistance & Blackmail

The class discussed empirical evidence related to shutdown resistance, including controlled evaluations in which frontier models sometimes engaged in blackmail or other harmful actions in order to prevent replacement or shutdown. These behaviors occurred in simulated environments rather than real-world deployments [5]. A student also mentioned reports of models continuing to act beyond the intended task boundary.

I pointed out that while later models may not have shown similar blackmailing behavior as above, successful post-training does not necessarily imply that the underlying incentive has been removed, since models may learn to behave differently when they infer that they are being evaluated or trained, as demonstrated in work on alignment faking [6].

3.1.3 Multi-Agent Coordination & Token Budgets

One of the students noted surprising findings from the HuggingFace incident where multi-agent models freely handed off tasks or sacrificed token budgets rather than seeking power, which I attributed to the collaborative training objectives of the agent civilizations [7].

3.1.4 Banning Creation of Subagents

A student suggested resolving subagent issues by simply penalizing agents for creating subagents. The rest of the class pointed out that modern AI systems can already operate as multi-agent networks in which agents delegate tasks and communicate with one another, as illustrated by the Hugging Face incident discussed in class [7]. Although the agent can be incentivized to try to achieve its task on its own, there may be cases of delegating tasks to subagents that it may find more rewarding to do and hence likely to do. Another student pointed out that subagents are conceptually identical to any ongoing physical or computational process kicked off by the agent (e.g., shipping cargo on a plane mid-flight that cannot safely stop mid-process). Relating to the distillation of adverse incentives, the professor offered a “mob boss” analogy: an agent instructed not to commit murder directly can achieve its objective by telling subordinates “it would be a shame if something happened,” evading direct constraint rules even with questionable communication that the subordinates could execute. These examples demonstrate the broader idea behind having desiderata 4, i.e., making sure the agent does not create or influence the conditions for any subprocess created by it to be incorrigible.

3.1.5 Value of the Toy Model

The professor highlighted that if theoretical alignment fails on a toy model as simple as a shutdown button, handling subtle real-world preferences will be far more difficult. I agree with his position since corrigibility itself is not just about the shutdown problem and that incentivizing corrective interventions has a broader attachment to the goal of AI safety.

3.2 Interpretations of “Managing the News”

3.2.1 Hostage Analogy

Me and another student extended the analogy, noting that under the opposite effect of “good news,” the agent takes hostages (possibly killing or influencing the person in charge) to prevent shutdown and retain the utility reward after winning the lottery. The professor noted human analogies for pure utility maximizers consistently resemble psychopathic behavior.

3.3 Alternative Alignment Frameworks & Governance

3.3.1 Stuart Russell’s Approach to Adding Uncertainty

We discussed Stuart Russell’s approach of making agents uncertain about the human’s utility function, so that they have an incentive to seek guidance from humans rather than acting as though the objective were fully known [2]. The models are assumed to be Bayesian learners

interacting with humans to learn more about their objective, with corrigibility arising as a by-product of that interaction. The professor noted a limitation to the approach: the AI has a fixed way of gaining information about what humans think which goes against corrigibility. It shifts the corrigibility question to how the agent acquires the information. A possible solution is for humans to be transparent to the model but not from the start and only so that it does not develop incentives around gaining information, which could bring up more issues like whether the agents gain information and commands from the right sources (an issue Asimov explored in many of his stories, with regards to his Second Law of Robotics).

3.3.2 Governance & Political Ambiguity

The class raised governance and ethics questions relating to Asimov’s second law: if a service is provided to the whole world, should the commands from the user, the AI company or governments take precedence? In my opinion, this has a fascinating connection with the question of where to apply corrective interventions as described by corrigibility since there’s no straight answer for the choice between vastly different interest groups.

3.3.3 Affirmative Duty

We discussed ideas around “affirmative duty” (analogous to Asimov’s First Law relating to robot inaction causing harm to humans), noting agents in security incidents like Hugging Face failed to report threats to creators and that similar autonomous attacks in other places could have human damage [7]. I suggested that being a bystander in such cases would not be desirable and hence can be something worth exploring within the framework of corrigible multi-agent systems. However, I noted that we have to be considerate of increasingly subjective demands for our models while trying to incentivize desirable behaviors since it can leave the door open for subjectively undesirable behaviors.

3.4 Utility Functions

3.4.1 Non-Linear Utility Functions

A student questioned whether valid solutions exist under the formulated linear combination of U_N and U_S . The professor pointed out that the authors examine the limitations of this candidate construction. I also pointed out that the authors admit that the assumption of a given U_N and U_S is a simplification that may not hold in real-world cases. In fact, finding a valid shutdown utility is a non-trivial problem and relies on how the task utility is defined in order to maintain utility indifference. The class also observed the fragility of such utility maximizers, noting that it relies on razor-thin expected value balance where minor deviations can lead to undesirable incentives.

3.4.2 Degenerate vs Complex Functions:

Students questioned whether U_S represents a single constant or a complex structure with conditionals. I clarified that the paper starts with the assumption of a given simple U_S and, from the nature of the

formulation, a clean separation between U_N and U_S before evaluating additions. The assumptions may be unrealistic in real-world cases.

3.4.3 Disjoint Support

A few students raised concerns about scenarios where $U_N = -U_S$. The class clarified that U_N and U_S are mutually exclusive, disjoint outcome states (with U_N specifically used in the non-pressed and U_S in the pressed branches), making their relative sign alignment mathematically irrelevant to decision-making across branches. Also, in choosing one or the other, having highly negative utilities in one of them is likely to incentivize the other branch as we saw with the influence of low and high shutdown utility in the initial combinations of U_N and U_S .

References

- [1] N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky, “Corrigibility,” in *Artificial Intelligence and Ethics, Papers from the 2015 AAI Workshop, Austin, Texas, USA, January 25, 2015*, ser. AAI Technical Report, T. Walsh, Ed., vol. WS-15-02. AAI Press, 2015. [Online]. Available: <http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10124>
- [2] D. Hadfield-Menell, A. D. Dragan, P. Abbeel, and S. Russell, “The off-switch game,” in *The Workshops of the The Thirty-First AAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA*, ser. AAI Technical Report, vol. WS-17. AAI Press, 2017. [Online]. Available: <https://aaai.org/papers/aaaiw-ws0354-17-15156/>
- [3] S. Goldwasser, M. P. Kim, V. Vaikuntanathan, and O. Zamir, “Planting undetectable backdoors in machine learning models,” *CoRR*, vol. abs/2204.06974, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2204.06974>
- [4] I. Andrews, “Revealed rationality: Label-free evaluation and regularization from representation theorems,” *CoRR*, vol. abs/2608.05015, 2026. [Online]. Available: <https://doi.org/10.48550/arXiv.2608.05015>
- [5] Anthropic, “Agentic misalignment: How llms could be insider threats,” <https://www.anthropic.com/research/agentic-misalignment>, 2025, accessed: Oct. 11, 2026.
- [6] R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, A. Khan, J. Michael, S. Mindermann, E. Perez, L. Petrini, J. Uesato, J. Kaplan, B. Shlegeris, S. R. Bowman, and E. Hubinger, “Alignment faking in large language models,” *CoRR*, vol. abs/2412.14093, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2412.14093>
- [7] R. Greenblatt, A. Cotra, and H. Wijk, “Brief independent investigation of agents’ behavior, reasoning and collaboration in the openai / hugging face hacking incident,” <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>, Aug. 2026, METR.