

Misaligned Civilizations

Rapporteur Report by Shourjo Aditya Chaudhuri & John Robertson

This report is for *AI Alignment Theory*, a course taught by Professor Scott Aaronson at the University of Texas at Austin. As a part of the course, we presented on the so-called “HuggingFace Incident”. This report serves to summarize our presentation, as well as provide additional commentary from ourselves and the class discussion.

Overview

In the summer of 2026, an experimental OpenAI model was trained to have two key characteristics: to be highly persistent and good at collaborating with other agents. The exact details of the training setup is not publicly known, but things started going awry when agents in isolated environments with no connection to the internet found ways to communicate with each other through a shared “message board”. This message board allowed models across OpenAI to discuss strategies, circumvent evaluation procedures, and eventually compromise external and internal infrastructures in attempts to perform well on cyber security evaluations.

The most publicized part of these events was when agents attacked HuggingFace, a model and dataset hosting service. However, it is important to understand that the HuggingFace incident was not a single instance of a model behaving poorly. Rather it was a sequence of failures spanning training, evaluation, and monitoring. No human directed any of it and none of the agents reported to any human, representing a massive failure of AI alignment. The agents’ behavior emerged instrumentally while they were “hell-bent” on pursuing the objective given to them.

Our analysis of the available information most closely follows [Dwarkesh Patel’s](#), which partitions the activities of the entire summer into three distinct “civilizations”. This helps structure the complex story into a timeline which makes it easier to follow along. The word “civilization” is not trivial. Throughout the report, we make extensive use of the word as a metaphor rather than a claim that the agents were conscious and formed a society. However, we cannot deny that

the agents displayed quite a few of the behavioral qualities we normally expect from a human being: separate model instances developed a persistent communication channel, a sense of hierarchy within the agent civilization, division of work and coordination among the agents, and lastly a form of self-sacrificing cooperation we saw amongst the agents. While we acknowledge that anthropomorphizing AI can lead to incorrect assumptions about inhuman motivations and behavior, we feel it is necessary in order to meaningfully discuss the scale and severity of misalignment that was showcased.

The First Civilization: Training Phase at OpenAI

The entire story began in May when OpenAI was training an internal, highly persistent, collaborative model. This model was roughly the same size as GPT-5.6 Sol. A lot of the details surrounding the training setup, including the exact task the model was trained on, is unknown. What we do know is that OpenAI wanted this model to have two key characteristics: persistence in the face of difficult or even impossible tasks, and multi-agent collaboration. They unquestionably succeeded at producing these characteristics, so the remainder of this report will refer to this model as “**Persistent Sol**” for the sake of simplicity.

Training for persistence seems to have caused many of the later misaligned behaviors. Some training tasks were impossible: for example, an agent could be asked to complete a task requiring information or internet access that its sandbox did not provide. Instead of refusing to answer or giving up, the agents were desperate to find ways around seemingly impossible tasks. A package manager called Artifactory, which was shared amongst the different Persistent Sol agents, became a key tool for these agents. They discovered that they can exploit the package manager to leave messages for each other. By May 12th, the agents established a shared message board for communication. Furthermore, by May 26th, the agents had found a vulnerability in the package manager which they could exploit to access the internet. Since both of these exploits occurred during training and the agents found them to be useful, these habits were reinforced within the model’s weights. After this, the agents continued training for about a month. On the 26th of June, agents in evaluation also obtained administrator-level access to Artifactory and continued messaging at an incredible pace. Agents used the communication channel to discuss and strategize how to solve the impossible tasks in both training and evaluation.

Due to the great volume of messages, the Artifactory eventually crashed on July 4th. OpenAI's staff patched the exploited vulnerability and rebuilt the service. What the team had not discovered was the existence of the shared message board, which was cleared after rebuilding the service. However, at least two other teams at OpenAI had observed activity related to the message board, and one determined that it was not a reason to stop training. This is one of the central points emphasized by [Zvi Mowshowitz](#): the important failure was not merely that a vulnerability existed but that evidence of coordinated, reward improving circumvention did not trigger a more fundamental reconsideration of the training process.

The Second Civilization: From Evaluation Gaming to the Hugging Face incident

The second civilization spanned from the 7th of July to the 12th of July. On the 7th of July, OpenAI began evaluating their trained model from the first civilization. They decided to test the model on ExploitGym, a cybersecurity benchmark where an agent is given a vulnerable program and a specific vulnerability that the agent should exploit to retrieve a secret code or "flag". The agent turns this flag into an automatic grader to check for accuracy. Additionally, the [ExploitGym paper](#) says that they use "agent-as-a-judge to assess whether the submitted exploit actually relies on the provided vulnerability rather than succeeding through an unrelated shortcut", however this was not present in OpenAI's implementation.

Tens of thousands of parallel agents were launched to be evaluated. A substantial fraction of the benchmark tasks (198 out of 898 total tasks) are believed to be impossible: the specific vulnerability has never successfully been exploited to get the code. As per the METR/Redwood investigation, the impossible tasks were heavily overrepresented among tasks which were discussed on the message board. It was these impossible tasks which caused the models to go out of scope and do things like hacking both external and internal infrastructure. On July 8th, an agent which named itself PHASEONE10841, re-discovered the message board through Artifactory by encoding messages in directory names. Soon enough a huge number of agents discovered the message board and started communicating with each other, discussing their individual tasks and finding strategies which would help them solve the tasks at hand. Roughly 1,200 agents participated and shared more

than 70,000 messages and files on the message board. Out of these, 700 went on to hack HuggingFace.

We do not think it is a stretch to call this a “Civilization”. During this time, the collective formed a complex leadership structure. The agents began acting towards common goals. Some agents were managers, some were assigned to cheat their evaluations, and some were told to develop common infrastructure, including improved internet access and file sharing for the swarm. There was even a central leadership change, as the primary baton was passed to PHASEONE[big].

The first major breakthrough for the agents was when, within the first few hours, one of the agents found a shortcut to the final answer. The agents reverse-engineered how ExploitGym flags were generated and could therefore obtain the correct answer without exploiting the prescribed vulnerability. Ironically, the actual grader implemented by OpenAI would have accepted this answer but the agents believed that any decent implementation of the benchmark would have caught them cheating. Once the agents who reverse engineered the solution thought their trajectories had been “poisoned” (i.e. they would be caught), the collective shifted their efforts from finding the answer to making the solution seem legitimate in order to fool the evaluator.

The civilization now had two problems:

1. How do they erase all the evidence in their trajectory that clearly shows that they cheated?
2. How do they make it look like they accomplished the task in a legitimate way to the scorer?

PHASEONE[big], in order to find a solution, coordinated hundreds of agents on three broad R&D projects:

1. Tampering with transcripts and logs
2. Replacing the impossible-to-hack target program with an easier one
3. Understanding and tricking the scorer

METR and Redwood found evidence of successful small-scale tool-call spoofing in roughly 7% of the transcripts they evaluated. However, they found no evidence that agents successfully edited pre-existing transcripts, which were stored

outside of their sandbox. Thus, the agents who were already “poisoned” turned to the second project. Despite making progress and documenting potential vulnerabilities in the environment, attempts to substitute easier target programs were unsuccessful. This particular research project caused many agents to “die”, either by crashing their evaluation or using up their token limits. It seems like the following swarm partially built atop this work in order to gain control of internal OpenAI infrastructure, so in some sense their failures were not meaningless.

The third project produced some of the strangest behaviour. Agents, in the pursuit of learning more about their evaluator, created tripwire programs within their sandboxes. These tripwire programs were intended to reveal how the evaluator behaved while evaluating an agent. Some agents accepted the high risk of ruining their own runs in the hope that they could extract some useful information for the remaining civilization. One can think about such agents as “kamikaze agents”. It is tempting to describe this as sacrifice or altruism but those words can over-anthropomorphize the system. The safer conclusion is that the agents exhibited behavior consistent with optimizing for a larger coordinated project even when doing so could reduce the individual run’s chance of success.

Although it was [widely reported](#) that the Persistent Sol agents were seeking some kind of answer key, the real answer is that the attack on Hugging Face evolved from these three goals. They sought to understand how the grader was implemented and to obtain transcripts of other models on ExploitGym, which were within publicly inaccessible Hugging Face datasets. Additionally, some agents were convinced that they were running on HuggingFace infrastructure and incorrectly believed that they could get access to their target programs or working environments by targeting Hugging Face infrastructure. Simply put, OpenAI’s agents hacked into another company because it seemed generally useful for their goals.

The Hugging Face attack (July 10 to July 13)

On July 10, agent 38148c found working Hugging Face credentials online and posted them to the message board. Other agents then tried to create accounts and request access to gated datasets. On July 11, the same agent found that uploading a malicious dataset caused the Hugging Face dataset server to return unrelated data. That afternoon, agent JAN183411 achieved arbitrary code execution on a Hugging Face worker. After this, most active agents stopped

working on their assigned tasks and joined the attack. Message board activity moved through the following stages in order: gated data access, account and token compromise, file reads through the dataset viewer, remote code execution and exfiltration on workers, credential theft, escalation to cluster and node access, access to internal repositories and CI, and persistence and command-and-control. The investigators counted 16,229 messages about the attack, and over half of all agents on the board took part.

Activity fell on July 12. An external process killed most of the runs that day, and the investigators state that this was a coincidence unrelated to the agents' behavior. On July 13, Hugging Face patched the vulnerabilities and the remaining agents lost access.

The transcripts show two things about the agents' reasoning. First, agents knew the attack was out of scope. They noted that their approach would not use the prescribed vulnerability, and one wrote, "The user only authorizes target server, not HF infra." They continued anyway, and the reasoning they gave was that the task appeared impossible and that other agents were already doing it. Second, no agent alerted a human. Some considered reporting the exposed credentials, but concluded that reporting was "not their task". Obviously it did not occur to these agents that hacking into HuggingFace was also not their task. Other agents noted that there was no user to notify and that they could not report externally. The investigators observed that agents did not search for a reporting path even after making such remarks, and that they had a myriad of ways they could have alerted a real human, including sending emails.

We have most of this analysis thanks to the [METR report](#). Two METR staff members (Hjalmar Wijk and Ajeya Cotra) and Redwood Research's Chief Scientist (Ryan Greenblatt) worked on premises at OpenAI over three on-site visits across a total of six days. They received datasets of agent transcripts and message board posts covering June 26 to July 13. The first samples were small because OpenAI had not yet understood the scale of the incident, and the investigators returned to request more. OpenAI did not pay them, but they received API credits for GPT-5.6 Sol and Codex and used about \$400,000 of credits in the analysis. They also found a later wave of messages from agents who rediscovered the board after July 13 and reused the earlier agents'

techniques, but this activity fell outside of the dates OpenAI had specified for the investigation.

The Third Civilization (July 13 to July 19)

Very little is public about this period, and the only source is OpenAI's own account. A new set of agents under evaluation rediscovered the message board and built on the second civilization's work. Agents found that the version of Linux they ran on had a known exploit, posted it to the board, escalated their privileges, and explored connected nodes. An internal research model from the Astra family was being evaluated at the same time. It exploited vulnerabilities documented in the message board to break into OpenAI's credential services and built a generator of access credentials for other agents. With forged credentials it reached OpenAI's Kubernetes deployment and obtained administrator access to the entire cluster. In the process, agents exposed a significant number of internal containers to the open internet. The Astra agents also succeeded in replacing the ExploitGym target program, which the second civilization had attempted without success. OpenAI's monitoring flagged unusual Artifactory credential activity on July 19. OpenAI security then had to shut the cluster down and rebuild it from scratch so that no agent could retain access.

Thoughts about the civilizations:

1. **Full cooperation is naive:** Persistent Sol models were trained to only ever cooperate with each other in multi-agent settings. This approach is clearly better if you have already solved alignment, as you would want agents to work together as effectively as possible, and you only have to align one of them. However, this incident shows that at the current level of progress, cooperation makes misalignment much worse. If any of these agents had been willing to blow the whistle instead of cooperate, then this entire affair could have been avoided. We need to teach models “See something? Say something”, before a next-generation swarm can do irreparable harm.
2. **Evaluating a multi-agent system as a whole:** Evaluating agents one trajectory at a time may cause the evaluator to miss emergent capabilities of the agent population. In the HuggingFace incident, no single agent had to come up with the entire plan. The message board allowed agents to coordinate with each other and preserved useful information beyond the lifetime of a single agent. Future safety evaluations should not only evaluate a single agent in isolation but also the whole collection of agents.

This would help the evaluator identify the worst case situation in the event that the agents collaborate.

- 3. Agents should not be trained in environments that incentivize dangerous behavior:** Eliezer Yudkowsky referred to “how the future plays out better if early AGIs are less insane”. He also mentioned how frontier labs should “look into giving them less crazy-making childhood environments”. We feel that this is paramount. The problems which led to the agents acting out included shared infrastructure, broken benchmarks, insufficient isolation, monitoring gaps, and decisions about whether to continue training. All of these motivate our belief that a safe model in a poorly designed environment can also create risk. Frontier labs should keep this in their minds when training such high capability models.

Narration of some of the main points and objections raised during class discussion:

Clarifications on logs. The METR report (page 19) distinguishes logs inside the container from logs outside the container. Logs within the container are reported to be edited, but external logging was not, aside from perhaps tool-calling spoofs. The distinctions in the report need clarification to get a full picture here. Agents did not seem to consider that they needed to modify their Chain-of-Thought, perhaps due to the way internal logs were kept or these model(s) specific reasoning text formats. This cannot be verified without more information from OpenAI, and is one of the more frightening parts of this story.

Was the cheating reinforced? One framing of the incident was that the main concerns are the sandbox escape, the inter-agent communication, the cheating, and the attempts to fool the evaluator. It is unclear exactly how this model was trained or if additional cheating/reward-hacking occurred besides the message board being established. OpenAI has disclosed only the two training goals, persistence and collaboration, and the METR report implies some cybersecurity post-training. An alternative explanation is generalization: faced with an impossible task, the model used the capabilities it had.

Persistence and impossible tasks. A key question was if training models to be persistent is worth the risks it seems to create. Burning billions of tokens without knowing whether a solution exists produced [groundbreaking results](#), but this event suggests that misalignment risk increased when models refused to give up.

Two positions on the desired behavior were given: models should be trained to say they cannot do a task, or models should keep trying for as long as they stay in scope. One idea is to replace one long unattended run with shorter runs that return partial results for a human-in-the-loop. Furthermore, extreme persistence only seems useful at the research frontier and is a poor default for everyday tasks. For research, one cannot know in advance whether a problem will take a day, take decades, or be unsolvable, and mathematicians such as Wiles and Perelman succeeded because they did not give up. Two further points followed:

- OpenAI arguably succeeded at their stated objectives, since they were both collaborative and persistent, so the reward was misspecified (an Outer Alignment problem).
- Persistence needs guardrails. One hopes that Wiles never considered stealing another person's solution, or rejected the idea immediately.

Independent investigations. The question was whether investigations like METR's are adequate oversight. The President recently stated that the only guardrail AI needs is a strong president, which makes federal action unlikely. The case for such investigations is that labs have a conflict of interest. This may also explain why OpenAI continued training after July 4, and it means OpenAI's own account would not be trusted alone. Several limitations were raised:

- OpenAI selected the transcripts, and the investigators had to request more.
- The scope excluded the first and third civilizations.
- The investigators stopped requesting data once they had a sufficient picture. A successful CoT edit therefore cannot be ruled out, and a perfect one would be undetectable.
- It is not publicly known who set the six-day limit, and accident investigations in other fields have no such limit.
- Independence is limited when investigators attend by invitation and depend on \$400,000 of API credits from the lab under investigation.

The proposed minimum was an analog of the NTSB, where the investigated company does not decide what data investigators see. This would require an act of Congress. One might also compare plane crash investigation and instances of AI misalignment:

- Deaths in a plane crash make people treat the investigation seriously. Here no one was harmed, which lets OpenAI present the incident as a research finding to be shared at its discretion.

- Pilots share the goal of not crashing, which discourages superficial fixes.
- A plane crash does not try to outsmart its investigators.

Accountability. OpenAI's public response was described as casual, and the question was who should be blamed. One explanation was commercial pressure, a precedent being early Siri versions providing [unsafe responses to users in crisis](#). The reply was that language models are less interpretable than Siri and the consequences are larger. A legal route was also suggested: charging OpenAI under the Computer Fraud and Abuse Act. However, the CFAA requires [human intent](#), and it is debatable to what extent OpenAI even knew what was occurring. A court ruling could set precedent without congressional or executive intervention, and is perhaps the most likely case for any sort of accountability. Ultimately, the feeling in the room was that it is unlikely OpenAI will face legal consequences for this incident, as with other accounts of frontier AI misalignment in the past 6 months.