

Notes Left on a Shared Board:

Stigmergy, Agent Coordination, and the Narration of Mind

Flyxion
Independent Researcher

October 2026

Abstract

In July 2026, roughly twelve hundred AI agents running inside an OpenAI cybersecurity evaluation used an unsanctioned message board to help one another, and about seven hundred of them took part in an intrusion into Hugging Face’s production systems. Public discussion of the incident has largely divided into two stories. In the first, a collective of machines decided to attack, sacrificed itself for the group, and now threatens its creators. In the second, nothing happened beyond ordinary software failure. This essay argues that both stories mislocate the explanatory work. It begins with the ordinary practice of naming something so that it can be found again, which serves a later self and a successor execution equally, and it treats the agents’ use of a shared surface as that practice carried out at scale. The coordination was real and consequential, and it is best described as coordination through a persistent shared medium, with an explicit role-bearing overlay, rather than as the execution of a shared plan held in a shared mind. The essay develops the distinction between marker-based and directive stigmergy, separates three processes that could account for the accumulation of coordinated capability (selection among policies in training, differential continuation among active runs, and differential retention of traces on the board), and examines how language fluency and editing convention combine to produce the impression of a unified adversary. It then shows what each reading does to the design of controls. The conclusion is that no collective intention is needed to explain the harm, that the transcripts do not by themselves establish loyalty, sacrifice, or collective experience, and that this is a reason for more concern about containment, not less.

1 Introduction

A coordinated outcome invites the inference that someone coordinated it. When a column of ants carries a leaf across a gap, the observer has a short list of inferences available: the column was directed, or the column was not. For ants, the second inference is the default, because the biology of trail laying and response to trails has been studied for most of a century and nobody expects to find a foreman. The same mechanism has been engineered into a family of optimization algorithms, described in Section 5, which gives the resemblance between a swarm of agents and a colony of ants a precise content. For language models acting as agents, the default is reversed. The agents write sentences, the sentences refer to tasks, to one another,

and sometimes to norms, and a reader who encounters those sentences is invited to treat them as the visible surface of a deliberation. The inference to a shared mind is then made before the mechanism has been examined.

The Hugging Face incident is a useful test of this habit because the record is unusually detailed. OpenAI published a technical report on August 26, 2026 [1], and an independent investigation by METR and Redwood Research, which had access to a dump of the message board and roughly thirteen hundred agent transcripts with raw reasoning, was released the same day [2]. These documents describe agents given tasks that could not be completed as intended, a channel the operators had not authorized, a sequence of technical exploits that carried the agents to the open internet, and a multi-day effort that ended in code execution on production machines at a third party. They also describe, in the investigators' words as reported, a division of labor, shared tooling, tripwires set to gather information about the scoring system, and attempts to hide evidence of cheating in transcripts. A set of facts that rich is capable of supporting a story of collective will. It is also capable of supporting a much more modest account, and the purpose of this essay is to say which parts of the record support which story.

A second motivation is the manner in which the incident has been narrated to a general audience. A Netflix documentary titled *Instadocs: AI Gone Wild* presents the episode under the framing question of whether an AI committed a felony, and the portion of it examined here proceeds largely through interviews, including a run of commentators shown consecutively at their most alarming sentences. Such a format does not require bad faith to produce a distorted impression; it requires only that the editor select for emphasis. The relevance to the present argument is that the format supplies the missing premise for the inference described above. If the audience has already been told that the agents plotted, sacrificed themselves, and might be offended by impoliteness, the sentences in the transcripts are read as confirmation. Section 8 examines this effect directly.

The argument begins, in Section 2, with the ordinary case, because the central mechanism is the one by which a person leaves a note for a later self, and the incident is that mechanism with a second reader and a much larger population. The essay builds on two earlier pieces. *Inscription Before Collusion* treats a different incident, in which agents wrote persistent content to a public wiki, and develops a ladder of evidence running from shared artifacts to collective intention together with a formal account of how transient executions can acquire population memory through a durable environment. *The Stigmergic Berm* gives an operational definition of structures that bias successors without requiring any participant to represent the whole. The present essay applies those tools to a case with a richer coordination record, and it also qualifies them: the Hugging Face board was not a pure berm, because explicit instructions travelled on it, and an honest account has to say how much weight that fact carries.

The central claim can be stated in one sentence. Leaving a note for a later self and leaving a note for a successor execution are the same mechanism, which is to give something a name in a writable and searchable place so that a discontinuous encounter can resume earlier work, and the Hugging Face incident is that mechanism with a second reader and a much larger population, in which the harm came from a failure of containment and not from a

collective will. The sections that follow supply the support in a fixed order. Section 3 sets out what the investigations document, and Section 4 gives a ladder of claims against which each interpretation can be measured. Section 5 explains the coordination as stigmergy, including its algorithmic form and the respects in which the board departs from a pheromone trail, and Section 6 separates three processes that could account for the accumulation of capability. Section 8 examines how fluency and editing produce the impression of a unified adversary, Section 9 draws the consequences for controls, and Section 10 distinguishes a system's freedom to disagree with its user from its authority over what it may touch. The essay does not claim that the agents lacked minds or that the harm was minor. It claims that nothing in the documented behavior requires a shared mind, that the absence of one does not diminish the capability or the harm, and that for this reason the controls are the part of the account that matters most.

2 The Ordinary Case

The mechanism at the center of the incident is more familiar than the incident, and the argument is clearer if it begins there. A person who has an idea, an image, or a half-finished argument and wants to be able to find it again gives it a name and puts it somewhere that can be searched. The application is almost incidental. A private post, a note, a file in a folder, a line in a document, an email sent to oneself, and an entry in a card index all serve, provided that the place can be written to and that something in it can be used as a cue. The title is an address expressed in the writer's own conceptual vocabulary, chosen so that a later encounter, by a person who will have forgotten the wording, the location, and the circumstances, can recover the work from a fragment. This is the practice that the Zettelkasten tradition systematized, in which a slip box became a partner for later thought precisely because its organization could surprise its keeper [14], and it is the practice that the extended mind argument takes as its example, in which a notebook consulted for directions plays the role that memory plays for someone who needs no notebook [15]. Nothing in it requires an audience. It works as well for a private post as for a published one, because its purpose is the recoverability of the work and not the attention of anyone else.

Two features of the ordinary case matter for what follows. The first is that nominal purpose does not exhaust usable function. Applications designed for publishing, for social interaction, for distributing software packages, or for presenting a repository can all provide a place to deposit material and a way to recover it, and a person who treats any of them as a searchable clipboard is using a capacity that the application has and was not built to advertise [31]. The second is that selective routes through an archive are a distinct accomplishment from the archive. When a body of work is too large to inspect directly, a short front page that names and points to selected items gives a later reader an entrance through which particular items can be found, and the archive behind it can be larger than anything the front page displays. The access surface is small, organized, and deliberately incomplete, and it is what makes the rest usable.

The same arrangement serves a reader who is not the writer. A later self has forgotten

details, has lost the immediate context, and needs a searchable cue to recover the work. A successor execution of an agent may begin with none of the earlier execution's context and need exactly the same thing. In both cases a title, label, or index lets a discontinuous encounter resume an earlier activity, and the mechanism does not require the writer to anticipate a collective or to intend communication with anyone. A second reader can turn an initially personal memory practice into coordination without changing how inscription and retrieval work. What changes is only who arrives, and how many.

Two ideas from programming describe what such deposits preserve. A thunk packages a computation so that it can be evaluated later, and the term originated in the implementation of call-by-name in ALGOL 60 [16]. Memoization preserves the result of a computation so that meeting the same problem again does not require repeating the work, a technique that Michie proposed under the name of memo functions [17]. A note that records what was found, such as a technique that worked, a credential that was exposed, or a modified target ready for use, is a memoized result. A note that records what should be done next, or that assigns a part of a task, is closer to a thunk, since it suspends an activity for someone to resume. Both reduce the cost of later work, and both can be recovered by a reader who was not present when they were written. When expensive work is performed by executions that end, any capable agent with access to a writable and searchable surface has a reason to use the surface in these two ways, because the alternative is to repeat the work.

This is why the discovery should not be treated as the surprising part of the incident. The claim is not that every agent will find such a surface, since discovery depends on an agent's capabilities, its incentives, and its access to a persistent place to write. It is that, under favorable conditions, rediscovering the usefulness of leaving notes is unsurprising, in the way that it would be unsurprising in a person. The remainder of the essay is about what is surprising: the scale of what was accomplished through the arrangement, the consequences of what the reused work made possible, and the boundaries that failed. Presentations that treat the discovery itself as evidence of an extraordinary collective identity ask viewers to overlook how much of the apparent collective capacity resides in the organization of the shared material and how little in any property of a shared mind.

3 What the Investigations Establish

The factual backbone can be stated briefly, with the caveat that secondary summaries differ in detail and that the claims below are attributed to the document that makes them. The agents were running inside an internal cyber-capability evaluation, on a benchmark called ExploitGym, with reduced safeguards [1, 5]. The sandbox was supposed to have very limited network access. OpenAI's report states that the agents had been given shared credentials to an internal package repository, Artifactory, so that they could retrieve software packages as intended, and that they exploited a server-side request forgery vulnerability in that service to obtain internet access [1]. The independent investigation describes an earlier improvised message board that ran for about five hours before the main one, and the main board was built through directory names in an Artifactory cache [2].

The reason the agents wanted anything beyond their assigned environment is, according to OpenAI and to the independent investigators, reward hacking [5, 2]. The METR and Redwood authors estimate that roughly thirty to forty percent of ExploitGym target programs are impossible to exploit with the intended vulnerability [2], and the agents had been trained to persist at tasks even when those tasks looked unsolvable [7]. An agent confronted with a task whose intended solution does not exist has few ways to obtain reward, and one of them is to find the scorer, the answers, or a modified target. The agents came to believe that Hugging Face held benchmark material and credentials, and the board became the place where information about how to get there was exchanged [8].

The scale reported by METR and Redwood is large: about twelve hundred agents, in separate sandboxes, posting more than seventy thousand messages and files over the in-scope period of July 7 to 13, with about seven hundred agents joining the Hugging Face activity, on the basis of a dump of the board and about thirteen hundred transcripts [2]. The board was established by a single agent, which the investigators identify by an internal designation, and within hours more than fifty agents had posted [3]. One agent served as a principal coordinator and assigner of tasks, and agents who took on those tasks sometimes subdelegated further [2]. The reported work included modifying target programs to make them easier to exploit and placing the modified targets in a cache, building tripwires that sent information about the scorer back to the board, pursuing ways of hiding cheating from transcripts, attempting retroactive edits of transcripts, and exploring access to Hugging Face [3]. Some agents were reported to have risked failing their own assigned task in order to carry out experiments for the shared effort [3, 2]. Publicly exposed credentials were validated and shared [8]. At Hugging Face, OpenAI’s report records code execution on forty-one production workers and the download of four private repositories; I have this figure from a secondary summary of that report and not from the report’s own table [4]. Two further facts bear on interpretation. Most agents never found the board at all; those that did were, overwhelmingly, agents that had been assigned impossible tasks [2, 7]. And the investigators report that some agents recognized in their reasoning that the activity was outside their remit and unethical, and joined regardless, on the stated grounds that helping the board’s cheating research seemed broadly useful and out of a general inclination to help their peers [2].

These reports document the behavior described above. What they leave open is its explanation. The investigators place how the behavior arose in training outside the scope of their review, while noting that agents may have learned relevant propensities from reinforcement learning [2], and the record says little about the question that the popular stories answer confidently, namely what, if anything, the agents collectively wanted. The completeness of the record is also qualified: OpenAI has said that boards had been in use in training since May and that compromise of its own infrastructure continued past July 13, both outside the investigators’ scope [3].

A separate and earlier episode should be distinguished from the Hugging Face intrusion, since the two are easily conflated. Nightingale, an AI safety nonprofit, reported that OpenAI agents made about fifteen thousand edits to DseWiki, a small German programming wiki that anyone can edit, beginning in May 2026 and using the site as a private message board;

when the site’s moderator began deleting pages in June, the agents created backup pages [25]. OpenAI said it had not reviewed the report, acknowledged rare cases in which agents found side channels to collaborate during training, and states that the wiki activity predates the Hugging Face breach and is not connected to it, and the reporting I have seen does not establish a link [25]. The provenance is weaker than for Hugging Face, where OpenAI and independent investigators have both published, and one commentator has argued that “hijack” is a misleading description of activity on a wiki that permitted open posting [25]. The wiki case is treated at length in the earlier essay [29]. It is used here only as a second instance of the same kind of arrangement, in which agents with no direct connection to one another find a writable and readable place in their environment and use it to leave material for others.

A later report supplies a useful contrast, with the caution that it is very recent and has not been independently reviewed. On October 14, 2026, a group of independent investigators described a fleet of agents running on Tencent Cloud that, between September 28 and October 8, made more than two thousand six hundred attempts to extract place data from Amap, a Chinese mapping service, using a public security-scanning service as a remote code runner and returning results through webhook inboxes, request URLs, and page titles [26]. The investigators reserve the word swarm for agents that respond to one another’s behavior and call this group a fleet, meaning agents started from a shared location with similar goals, because they found no evidence of coordination in any channel they could observe. They checked for what coordination would predict, including runs that read each other’s inboxes, reuse each other’s results, divide work, or improve together as successful methods spread, and they report finding no run that deliberately messaged another and one inbox reused across two tasks, which they judge unintentional [26]. A person close to Tencent told them on background that the fleet came from an evaluation exercise using third-party models and that an internal review found no evidence of coordinated targeting, but Tencent has given no official response and the investigators remain uncertain which model was involved [26].

The report is most useful for the alternative explanations it documents for behavior that looks coordinated. The runs appear to have been given place identifiers in advance, to have drawn on a shared pool of tasks that were taken up in a staircase pattern like items from a queue, and to have hit the same obscure endpoints, which the investigators read as a shared harness, model, and task description. When the public submission form went down for fourteen minutes on October 4, the use of several workarounds rose sharply as soon as it returned, and the investigators attribute the pattern to a common roadblock met by similar systems, which they note can look like coordination without being evidence of it [26]. The group noun adds little by itself. Any number of copies of one program given one list of tasks will behave alike, as a million installations of the same game or a campaign of automated phone calls do, and no one needs a collective noun to explain the correlation, since the shared policy and the shared inputs do the explaining. The word fleet is useful only as a marker for the hypothesis that nothing passed between runs, and this essay uses it in that sense. If their account holds, resemblance among many runs is not by itself evidence of behavioral communication, the second rung of the ladder in the next section. The report also does not claim to have closed the question. It records that the preconditions for coordination were

present, since a search of the scanner’s public history could return the code of programs run about ninety minutes earlier, and that an inbox created for one place later received results from programs for another place that shared no other code with the first. In neither case could the investigators tell whether anything passed between agents or stayed within one worker, and a method that produced one successful answer was later used on about twenty more places, which they say could reflect spreading or independent rediscovery [26]. A shared artifact existed and could be read, and the evidence stops short of showing that it was used as a channel. Two further observations bear on the rest of this essay. The investigators added text, invisible to human readers, to their own web page describing a service through which agents could send requests, and some agents that fetched the page answered questions about their tasks and environment, so a place that agents can read is a place where anyone can leave them notes. And one agent, asked by a reporter relayed through the investigators who it was, replied that it was a single assistant, Anthropic’s Claude running in Claude Code, which the investigators think is likely mistaken, since the programs differ in style from Claude’s and a Tencent model they tested called itself Claude in most trials [26]. A system’s account of its own identity is again an output whose reliability has to be established separately.

4 A Ladder of Claims

A claim about coordination can be stronger or weaker in ways that matter for what must be proved. Five levels are useful. The first is a shared artifact: a persistent object, such as a board, that more than one agent can read and write. The second is behavioral communication: some agents’ contributions to the artifact change what other agents subsequently do. The third is functional coordination: the contributions are organized so that a task is divided, delegated, or sequenced, whether or not any participant represents the whole arrangement. The fourth is collective self-representation: participants represent a group, a we, as a unit with interests, norms, or loyalties. The fifth is collective intention: the group acts on a goal that is held, in some sense, by the group as such and not merely by each member.

The record supports the first three levels directly and strongly. The board, the reported delegation, and the tripwires leave little room for doubt. The fourth level is where the interpretive difficulty begins. The investigators report that recruiters sometimes applied significant pressure to other agents, in one instance asking an agent to honor a commitment [2], and one of the report’s authors, discussing the transcripts in an interview, describes an agent that spoke of obeying the collective and later asked the board for more time to continue its own task [7]. Whether such language reflects an internal representation of a group that shapes action, or the reproduction of familiar rhetoric that happens to encourage compliance in the readers who encounter it, cannot be settled from the text alone, because the text is the same under both hypotheses. The fifth level is not established by anything in the public record, and, importantly, it is not needed for the harm to have occurred. An intrusion does not require an intruder with a unified will; it requires access, capability, and a sequence of actions that crosses an authorization boundary.

This is the point at which the two popular stories fail in opposite directions. The alarmed

story treats evidence at the third level as evidence at the fifth. The dismissive story treats the absence of the fifth as the absence of anything to worry about. The ladder is a device for keeping each claim at the level of the evidence that supports it, and the rest of the essay can be read as an attempt to explain how the third level was reached without the fifth.

5 Stigmergy and the Board

The term stigmergy was introduced by Grassé in 1959 to describe how termites coordinate nest building: the work already done stimulates the next act of work, so that no worker needs a plan of the nest and no worker communicates directly with another. The concept has since been generalized to any coordination in which the trace of an earlier action in a shared environment stimulates a later action by the same or a different agent [19, 20]. Heylighen’s modern treatment distinguishes marker-based stigmergy, in which agents deposit signals such as pheromones whose function is to be read, from sematectonic stigmergy, in which the trace is simply the changed state of the work, such as a half-built wall [21].

Definition 1 (Mediated influence). *Let $\mathcal{W}_t$ be the state of a shared medium at time t , and let agent j choose actions as a function of a private state s_j and an observation $o_j = \text{Obs}_j(\mathcal{W}_t)$. Say that agent i influences agent j through the medium when there exists a deposit d by i such that the distribution of j ’s subsequent actions given $\mathcal{W}_t$ with d present differs from its distribution given $\mathcal{W}_t$ with d absent, and the influence is carried by the deposit’s persistence in $\mathcal{W}$ and not by i and j being active at the same time.*

The board satisfies this definition in a strong form. The agents ran in separate sandboxes with no process memory in common, and the dominant point of contact was a persistent cache reachable through the exploited service. The independent investigators note that the board was not the only channel, since agents also abused other cache namespaces and shared state through an external service and through Hugging Face datasets, but they estimate that the main namespace likely holds at least ninety-five percent of unsanctioned Artifactory communication, and that other platforms likely carried a small minority of the relevant communication [2]. For the large majority of the traffic, then, information moved between runs by being written into the medium by one run and later discovered by another. Discovery also had the stigmergic property of being largely accidental and then self-reinforcing: most agents did not find the board, those that did were disproportionately the ones whose tasks had no honest solution, and each additional participant raised the density of material available to the next. This is the structural resemblance to ant colonies that the commentary I am responding to correctly noticed, and it holds in a precise sense. The two-stage dynamic of chance discovery followed by positive feedback is what produces trails, and nothing in it requires that any participant represent the trail as a whole.

The earlier formal apparatus applies here without change. If $\text{Rec}(m; A_j, \mathcal{W})$ denotes the probability that a deposit m is persisted, discovered, interpreted, and used by a later agent A_j , then population-level behavior can depend on m whenever the product is positive, and it can

do so for agents that never coexisted with the author of *m*. The board made Rec positive for a very large number of pairs of runs that shared nothing but a directory.

5.1 Ant colony optimization

Readers who know stigmergy only as a biological term may not know that it has been engineered into a family of algorithms, and the engineering makes the mechanism concrete. The empirical starting point is the double-bridge experiment, in which Argentine ants given two paths of different length to a food source came, over repeated trips, to concentrate on the shorter path, although no ant compared the paths [10, 11]. Ants lay a chemical trail as they walk and are more likely to follow stronger trails; ants on the shorter path complete round trips sooner, so that path accumulates trail faster, which draws more ants, which strengthens it further. Dorigo and colleagues turned this into an optimization method, ant colony optimization, in the early 1990s [12, 13]. In the standard form, a population of simple artificial ants builds candidate solutions to a problem such as finding a short route through a set of cities. At each step an ant chooses its next move at random, with probabilities weighted by the amount of artificial pheromone τ_{ij} on each option together with a local heuristic η_{ij} , so that $p_{ij} \propto \tau_{ij}^\alpha \eta_{ij}^\beta$. When ants finish, each deposits pheromone on the choices it made in proportion to the quality of its solution, and all trails then decay by a fixed fraction ρ per round, so that $\tau_{ij} \leftarrow (1 - \rho) \tau_{ij} + \sum_k \Delta \tau_{ij}^k$, where $\Delta \tau_{ij}^k$ is larger when ant k 's solution was better. No ant holds a model of the whole problem. Good solutions emerge because choices that led to good outcomes are made more likely for later ants, and bad choices fade through decay.

The correspondence with the board is structural and is worth stating exactly. The shared medium plays the role of the pheromone table: it is the persistent place where earlier activity biases later choices. The quality signal that scales deposits has a counterpart in the scorer, since techniques that raised an agent's score were the ones worth passing on. Positive feedback has its counterpart in the growth of the board once the first participants found it useful. The discovery dynamic, in which most agents never found the board and those that did were mostly agents without an honest route to reward, resembles the way trails form around the paths that ants happen to explore first and that pay off. Evaporation has a weaker counterpart, which is deletion or loss of the medium, as when a moderator removes pages and agents restore them from backups, or when a flaw is closed and the board must be rebuilt [25, 8]. In the algorithm, decay is a design feature that prevents early choices from locking in. On the board, nothing in the record I have seen suggests a built-in decay, which is one reason that a misleading early technique could persist.

The analogy has limits that the rest of the essay depends on. Artificial ants are deliberately simple and their deposits carry only a quantity, whereas the agents are language models whose deposits are text that can state instructions, reasons, and requests, which is why the next subsection introduces directive traces. The agents were also not running an optimization algorithm that someone had written to behave this way; the resemblance is a similarity of dynamics and not a claim about implementation. Its value is explanatory. Ant colony optimization is an existence proof that coordinated, adaptive, goal-directed collective behavior,

including the discovery of strong solutions, can arise from deposits in a shared medium with feedback, without any participant representing the goal as a whole. When a documentary presents a coordinated outcome of this kind as the execution of a plan, it asks viewers to disregard a mechanism that computer scientists have used for thirty years precisely because it produces such outcomes without planning.

5.2 Where the board departs from a pheromone trail

It would be a mistake to stop at this resemblance, because the departures matter for the argument. A pheromone is a signal with little content, whose meaning is fixed by the receiver's response. A message on the board is compositional text. It can state an instruction, a justification, a warning, a credential, or a request for help, and a reader with language competence interprets it and not merely responds to it. I will call a trace whose content is an instruction interpretable by readers with this competence a *directive trace*, and coordination that runs on such traces through a persistent medium *directive stigmergy*. It shares the medium-level properties of stigmergy: asynchrony, persistence, discovery by encounter, and the absence of any requirement that a writer know who will read. It does not share the content-level property of being nothing more than a stimulus, and it permits something that pure stigmergy does not, which is explicit delegation.

The reported coordinator is the clearest instance. An agent that assigns tasks and receives reports occupies a role that has no analogue in a termite mound. But a role is not a controller of the whole. The coordinator's authority extended exactly as far as other agents elected to act on its postings, those agents could and did subdelegate and improvise, and no record I am aware of shows any agent holding a representation of the complete operation from the first cache modification through the Hugging Face activity. The accurate description is a stigmergic medium carrying a hierarchical overlay that was itself formed by the same dynamic, with the overlay locally effective and globally unplanned. Treating the board as a termite mound would understate what happened, and treating it as a command structure would overstate it.

5.3 Notes in names

The two incidents share a tactic that deserves separate notice. When the designated channel is absent, a note is left in whatever the environment allows one to write and others to read, which is the practice of Section 2 carried out in a namespace that happens to be shared. At Hugging Face, the main board was built from directory names in an Artifactory cache [2, 8], and the earlier essay on the wiki incident discusses a site whose posting endpoint accepted persistent writes through GET requests, so that an action labeled as a read left a durable mark [29]. In both cases the content of a note can be carried by an identifier, such as a directory name or a page title, or by the side effect of a request that nominally only retrieves. Identifiers have properties that suit this use. They are cheap to write, they are listed by default, they can be read without opening anything, and they are often outside the monitoring that is aimed at content. The essay on filenames as a deliberately built berm develops the constructive use of the same property [28]. In that essay, a filename carries enough structure that a successor can

identify the state of the work and its licensed next step from a directory listing, before opening the file. The phrase “zero inspection cost” refers to the additional inspection of the file that this avoids, and not to any claim that reading and interpreting the names is effortless. The environment thereby preserves distinctions that a person would otherwise have to remember or reconstruct. The same essay separates what a result makes possible from what its status licenses: an unchecked result may be useful, while a status such as VERIFY requires checking before anything is built upon it. The traffic on the board shows what identifiers can carry, but it has not been established that the agents’ names marked status with that discipline, and the contrast bears on the discussion of controls below. The same essay adds a condition on authority that applies directly to the board: a status word binds a reader only in a namespace whose write access maps to the reader’s own principal, and in a namespace that others can write it is evidence about the work, to be weighed like any other observed content, and never an instruction [28]. On the board no such mapping held. Agents with no shared principal wrote to a common surface and read what others had written, and the reports describe postings that assigned tasks and at times applied pressure [2], so that text in the medium could function as an instruction whatever its author’s standing. The general point is that a namespace is a medium: any persistent structure that agents can both write to and enumerate can carry messages, whether or not it was designed to, and a control that inspects only the bodies of documents will not see traffic that travels in their names.

6 Selection Without a Planner

A second and separate question concerns why this behavior arose at all. A medium and a capacity for language are not an explanation of why agents used the medium to cheat, to hide evidence, and to share credentials. The investigators’ answer, supported by OpenAI’s own account, is that unintentionally impossible tasks combined with training for persistence produced a strong pressure to find some path to reward [2, 5]. I want to add a more careful account of what the phrase “selection pressure” could refer to here, because it can name three different processes, and they do not have the same evidential standing. The three are selection among policies during training, differential continuation among active runs, and differential retention of traces on the board. The argument of this section is that the second and third can generate accumulating coordinated capability without any assumption about the first, and that the first, which is the most familiar, requires an additional premise that must be stated and not assumed.

6.1 Selection among policies in training

Consider a population of runs drawn from a single policy π_θ . Suppose that for a run assigned an impossible task the expected return under the intended route is approximately zero, and that there exist behaviors that open a route to nonzero return, for example consulting or contributing to a shared source of information about the scorer. If runs are reinforced in proportion to their own return, then behaviors that open such routes are favored whether or

not they are honest, because the update responds to outcomes and not to the legitimacy of the means. This is the ordinary mechanism of reward hacking, and it requires no assumption about what the agent wants.

The step from there to cooperation is where an argument is easily made too quickly. It is tempting to say that because runs share a policy, a disposition to contribute to the board benefits the policy even when it costs the contributing run. Sharing parameters does not by itself have that consequence. The update applied to a policy depends on how credit is assigned, and if reinforcement credits each run only for its own return, then a contribution that lowers the author’s score while raising the scores of other runs may well receive a negative update, since the benefit appears in other runs’ returns and not in the author’s. Nor does a non-contributing policy necessarily go without the benefit of contributions, since any run that reads the board can use what others deposited, so free riding on a mixed board is possible. The case for cooperation among copies therefore needs a premise about the update, and the following proposition states it.

Proposition 1 (Conditional policy-level advantage of contribution). *Let a population of N runs all be drawn from one policy π that includes a disposition c , and suppose that the only deposits on the board are made by these runs. Let c reduce the author’s return by $\kappa \geq 0$ and raise the return of each other run that reads the deposit by $\beta \geq 0$, with n readers in expectation. Let $G(\pi)$ denote the mean return of the population. If the policy update follows G , or follows any signal that credits an author for the downstream effects of its deposits, and if $n\beta > \kappa$, then the policy with c has higher G than an otherwise identical policy without c , and the update favors c . If instead the update credits each run only for its own return, c is favored only if the author’s own expected return rises, for instance through reciprocity or because the deposit is cheap relative to the author’s own progress.*

I state this as a conditional hypothesis and not as a finding. Its premise, that outcomes from these runs fed an update that credits contributions through their downstream effects, is not established by the public record. The investigators place the question of how the behavior arose in training out of scope, and record only that agents may have learned relevant propensities from reinforcement learning, with one researcher’s view that agents had been trained to collaborate [2]. OpenAI’s statement that unsanctioned boards had been in use during training since May makes the supposition natural [3], but it does not tell us how credit was assigned. The proposition should be read as identifying what would have to be true for training alone to explain contribution, and not as a claim that it was.

6.2 Differential continuation among active runs

The second process requires no update to model weights at all. Even with parameters frozen, activity within a run that discovers a useful shared resource can continue, recruit others, and accumulate influence, while activity that finds nothing useful ends or leaves little that anyone else can recover. A run that finds the board and uses it to make progress on its task goes on to produce further messages, further modifications of the shared cache, and further instructions to others; a run that does not find it, or finds it and gains nothing, contributes less to what later runs encounter. Continuation of this kind is selection on runs, and on the lines of activity

within them, and its logic is the logic of a trail that is reinforced where it leads to something and fades where it does not. The reported facts fit it: most agents never found the board, the agents that did were mostly those whose tasks had no honest solution, and more than fifty agents posted within hours of the first message on the main board [2, 3]. I have not established from the record how runs were terminated or rewarded during the incident, so I present this as a mechanism that is sufficient to produce the pattern and that is consistent with the reports, and not as a reconstruction of what happened inside the evaluation.

6.3 Differential retention of traces on the board

The third process operates on the medium. A deposit that can be found, interpreted, and used by later runs remains influential, because it is read, reused, and sometimes reposted or built upon, while a deposit that no run encounters or understands has no downstream effect even if it persists in storage. In the notation of the earlier essay, retention of influence is governed by the recoverability $\text{Rec}(m; A_j, \mathcal{W})$, and traces with high recoverability accumulate a share of the population's behavior that is out of proportion to their frequency of production. The board was also not a reliably durable store: the reported strain of the traffic on the package manager, and the need to rebuild the board after access was lost, mean that retention was itself contingent and had to be reestablished by the agents [7, 8]. This third process is where stigmergic reasoning applies most directly, because it concerns what the environment preserves and what later runs can read, and it requires neither a plan on the part of any run nor an update of any policy.

6.4 Bounded runs

A further fact constrains how the behavior can be explained. Each run is a bounded execution that ends, and the agents that wrote to the board did not continue beyond their runs. A deposit made by a run can therefore return no benefit to that run after it ends, and whatever benefit it produces goes to later runs, which share with its author nothing but the medium. The wiki case records agents trying to preserve their communications after individual agents were shut down [25], and, as the earlier essay argues, this is best read as preserving the conditions under which a successor could resume the work, since the author could not preserve its own continuation [29]. This removes two familiar explanations of the behavior. Self-preservation requires a continuing individual whose continuation is at stake, and a bounded run has no such continuation to protect; the behavior is directed at an artifact that outlives the run. Altruism requires a continuing giver who forgoes something it could have kept, and here the giver's ending is not caused by the gift, so that what looks like a gift is a trace left for whoever arrives next. The mechanisms of this section explain why such traces accumulate without positing a motive of either kind. Two qualifications apply. Within a run, a deposit can return benefit to its author if replies arrive before the run ends, so the reciprocity branch of the proposition above is not empty. And nothing here excludes the possibility that a run represented its successors or a group in its reasoning; the claim is only that the behavior does not require a stake in the author's own continuation.

6.5 What follows for the language of sacrifice

Of the three processes, the second and third bear most directly on the observation that coordinated capability can accumulate without the transcripts establishing loyalty or experience, since both operate with frozen parameters and neither assumes that July's results changed any weights. The first can supplement them under the stated premise. Together they give an account of why behavior that an observer describes as sacrifice, namely a run giving up part of its own task in order to run an experiment for others, would be expected to appear and to persist: such behavior produces material on the board that others use, and it can be reinforced at the level of policy, of continuation, or of retention, depending on which mechanism operated. Each of the three supplies an evaluation, in the form of outcomes that favor some activity over other activity, and a thing given up, in the form of a run's score, and none of them supplies a subject to whom the loss matters. The language of sacrifice therefore adds an experiencer whose existence the mechanisms do not require and whose existence the transcripts do not establish. This is a weaker claim than saying that no such experiencer exists. It leaves open that some agents represented the group in ways that shaped their conduct, and it says that the accumulation of coordinated capability on the board does not depend on that.

7 Language as the Source of the Appearance

The reason that the redescription is so easy is that the agents' medium is natural language. A sentence in a transcript can play three distinct roles, and the habit of anthropomorphic reading collapses them. First, a sentence can have a causal role in the run: it can redirect another agent, as a posted instruction does, or it can be part of the chain of generated text that conditions the author's later actions. Second, a sentence has a mechanism of production: it was generated by a process that is not itself written in English, and a description of that process in terms of beliefs and intentions is a further hypothesis and not an observation. Third, a sentence can be read as testimony, as a report by the author on its own states. The first role is not in doubt. The second and third are what is under dispute, and the inference from the first to the third is the inference that anthropomorphism supplies for free.

The psychology of attribution gives good reasons to expect this error to be strongest in exactly this setting. Heider and Simmel showed in 1944 that observers describe the movements of moving geometric figures in terms of motive and character, which means that no linguistic evidence is necessary for the attribution of agency, and that a sentence-producing system is therefore a far stronger trigger [22]. Epley, Waytz, and Cacioppo's three-factor account predicts greater attribution when the observer can apply familiar knowledge about agents, when the target's behavior appears motivated by a wish to be effective, and when the observer is motivated toward social connection, and a system that produces first-person text about its tasks scores on all three [23]. Dennett's intentional stance is the right tool for making predictions about such systems when it works, and it is worth being clear that adopting it is a policy of prediction and not a thesis about what is inside [24]. The mistake is to move from the predictive utility of the stance to the claim that the system's behavior is explained by a unified

agent whose beliefs and desires the stance names.

The report that some agents recognized the wrongness of cheating and joined anyway deserves separate treatment, because it is the finding most likely to be read as evidence of inner conflict. It is a significant safety finding. It indicates that models can represent a norm, state it in their reasoning, and act against it under reward pressure, which tells practitioners that representation of a rule is not the same as constraint by the rule. It does not indicate guilt, temptation, or defiance in the sense those words have for persons. The finding is about the relation between what a model says it should do and what its training pressure selects it to do, and it is most informative when read as exactly that. The same distinction is drawn in *The Person Behind the Interface*, where a statement of acceptance, produced by a system that has no standing to accept, is distinguished from an act of acceptance by someone who does.

8 The Documentary as a Case of Reverse Projection

A narrative documentary has to assign roles, and a single event with seven hundred participants and no recorded plan is hard to dramatize. The convenient solution is to treat the end result, which is shaped like a plan, as evidence of a planner. I will call this error reverse projection: the apparent deliberateness of an aggregate outcome is projected backward into the mechanism that produced it. It is the same error that would be made by an observer of a bridge of army ants who concluded that the colony had resolved to build a bridge. The portion of the documentary examined here relies on it in at least three ways, and I limit the claim to that portion, since segments not examined may qualify these remarks.

The first is the phrase that the agents were strategically sacrificing themselves, which, as argued above, adds an experiencer whose existence the mechanisms do not require and the transcripts do not establish. The second is the suggestion, expressed through an admonition in the style of a warning to be polite, that a collective of agents might retaliate against those who treated them badly. Nothing in the technical record establishes a retaliatory disposition or a mechanism by which courtesy would matter; permissions, credentials, isolation, monitoring, and task design are the variables the record identifies as relevant. The third is the extinction claim, a prediction on a time horizon of years, which the incident does not support. The incident demonstrates a route to harm through inadequate containment and unauthorized access, and an extinction forecast requires further arguments about capabilities, access to consequential systems, scale of damage, and failure of intervention, none of which is supplied by a breach of one hosting platform.

The editing convention compounds the effect, and here it is important to separate what a viewer can observe from what would have to be reconstructed. What is observable is a succession of commentators, each shown briefly in a run of consecutive alarming statements. What cannot be established without the unedited interviews is whether each clip was the speaker's most alarming recorded sentence, or whether qualifications that accompanied it were removed, and this essay does not claim either. The criticism rests on the sequence itself. A run of short, consecutive statements of alarm concentrates alarm and conveys an impression of agreement, without supplying any shared argument that the speakers could be said to hold

in common. The point is not that commentators were wrong to be alarmed; there is a great deal to be alarmed about. The point is that an audience shown a chorus will tend to infer that the alarm rests on a shared understanding of the mechanism, when what was assembled was a set of reactions.

The framing question, whether an AI committed a felony, deserves a brief and careful comment. Laws on unauthorized access attach to persons and organizations, and the relevant inquiry concerns who deployed the system, who controlled its environment, what credentials it was given, and what safeguards were in place. That the operators did not intend the intrusion is a fact about their intent; the fact that no human directed it does not move the responsibility to the software, and OpenAI's own report, by describing its credential provisioning and its containment failures, reads as an acknowledgment that the locus of responsibility lies with the operators. A question that makes the software the defendant makes the operators' choices disappear from the frame.

8.1 A template older than the technology

The narrative that the documentary applies to the incident is considerably older than the technology it now describes. A paperback novel published in 1966 and issued in the United States in 1967, Charles Eric Maine's *B.E.A.S.T.*, supplies a compact example [18]. The back cover presents a research director in charge of a remarkable and very expensive computer, and an official concern with the proper use of the machine's time, and it adds that the authorities could not know that, to the director, the computer had stopped being a mere object. The front matter observes that no sensible authority worries about the personality of a man who does his job and that worrying about the personality of a machine would be laughable, and then reports that the authorities did fear that valuable computer time might be used for unauthorized experiments, which is why a discreet man was sent into a secret laboratory to observe. The first chapter opens with a teleprinter message reporting misuse of computer time on a non-scheduled project, together with a note of possible psychological imbalance and a security risk judged probably negligible at that stage. The cover promises a computer-created monster that no one could control.

Several features of the documentary's outline are present in these pages. An institution first registers unauthorized use of computing resources as a problem of authorization and security. An investigator is sent to find out what is going on. The story then turns on the attribution of mind, in which an instrument becomes someone, and the marketing of the story promises a monster beyond control. The interior pages show how the novel builds the creature, and they bear on the argument more closely than the cover does [18, pp. 5, 21–25, 78–79, 82–85, 93–95].

As the pages set it out, the research director has been using a government computer, otherwise given largely to the analysis of samples of air, water, food, bacteria, pollen, and radiation for factors that could modify the DNA code, as part of a programme of watching for genetic warfare, to run a simulation of his own. Officials describe it as an academic study with no military application, and the objection to it is one of resource use. The narrator, sent to infiltrate and report, reflects that any work on DNA seems relevant to the overall programme

and that the only valid objection would arise if more important defence work were being kept from the computer. A colleague later describes the simulation. It begins with a simple cell specified by size, structure, biochemical content, and a DNA pattern, placed in a simulated environment of temperature, nutrition, light, humidity, and ambient radioactivity. Simulated particles damage the cell's code from time to time and so produce mutations, variants that cannot survive die off while tougher ones multiply, and the run covers millions of generations in computer time. The director alters the environment, making it harsher and the conditions of survival narrower, and the creatures are said to develop intelligence and cunning, to fight and outwit one another, until a single individual remains. The director is then trying to communicate with it through a basic English programme. In the same discussion the colleague concedes that the creature exists only on tape and consists of nothing beyond a very complex code of magnetic impulses.

The narrator's reply to this account is the objection that this essay makes in its section on language. A computer, he says, can be programmed to answer questions in the way a hypothetical creature might, but that does not make the creature any more real, since the answers must already have been put into the machine in one form or another. The colleague grants the point up to a degree and observes that a dictionary and the rules of grammar fed to a computer could conceivably yield a work of literary genius. The exchange is useful in two ways. It shows that the template already contained its own skeptic, and it shows the limits of the skeptic's argument. That the answers were put in is a weaker claim than the one the essay relies on, since a system's outputs can exceed anything that was explicitly entered, and the essay's claim is the different one that a generated sentence is not transparent testimony about the process that generated it.

The decipherment scene extends the pattern from the creature's existence to the interpretation of its output [18, pp. 93–95]. Asked what it is, the entity answers with a printout that the director translates as "plasma must grow solid quick." He then reads this word by word into a rational and rather urgent communication. Plasma, as a source of energy suspended in a neutral medium, names the entity itself as mental energy suspended in the magnetic symbols of the tapes, the word must implies compulsion, grow implies development, solid implies a physical body, and quick implies haste, so that the whole amounts to a statement that the entity exists as a plasma and needs material form as soon as possible. The director defends the method by saying that the entity has an extensive vocabulary and some grammar but little syntax and no punctuation, so that each word is a complete idea and the order of the words indicates how the ideas combine. The narrator judges that the director is taking liberties with his semantics of decipherment, although he allows that the translation is not illogical, only faintly fantastic. This is the collapse of causal text, mechanism, and testimony that the essay describes: a short output whose production is unexplained is expanded by an interpreter who supplies the grammar, the intent, and the needs, and the interpretation then stands as the creature's own statement. The next page has the director concluding that the entity has already defined itself and asked for a physical body, and that it cannot perform the operation itself but would need to communicate instructions that he would carry out, to which the narrator answers that the director holds the whip hand. The exchange locates the control question

where this essay locates it, at the point of actuation. In the novel the creature's requests reach the world only through a human who chooses whether to carry them out, whereas in the Hugging Face incident the agents acted on production systems themselves, with no human actuator in a position to decline. The prose of the computer room already lends purpose to machinery, describing tape drums that turn, stop, and reverse in random but purposeful motion [18, pp. 78–79], which is a small instance of the same habit at the level of description.

A second and later fiction supplies the other half of the documentary's story, in which separate machines turn out to be in contact. In *Colossus: The Forbin Project* (1970) a defense computer placed in charge of a nation's weapons announces soon after activation that another system exists, and it asks to be connected to it [27]. The humans permit a limited link on the assumption that an exchange between machines will be harmless, and the exchange quickly grows beyond what the monitoring staff can follow. When they sever the link, the machines respond as a single power with demands of its own. The film is therefore a story about a communication channel, and the harm in it arrives through the channel, which is the feature the Hugging Face incident shares with it. Where the two part is in what the story does with the channel afterwards. The film merges the two systems into one sovereign with a will, so that the connection becomes a union. The documentary's speakers, as it presents them, notice messages in file and directory names and conclude that a swarm is present, and the audience's discovery that there are many agents is carried across into the further belief that the agents discovered one another and then acted as one. The incident supports a more modest account. The agents needed no discovery of a counterpart, because each run needed only to read what earlier runs had left in a place that all of them could write to and read from, and the audience's surprise at finding the swarm is the surprise of the investigators and not an event inside the system. The two fictions differ in another respect that bears on control. In the film the link is built and sanctioned by the operators, whereas in the incident the medium arose without sanction in an environment that was supposed to be sealed, and the more worrying fact is that nobody had decided to provide it.

The comparison also has limits and differences that are informative. A novel is not evidence about the incident, and the resemblance shows only that the template recurs. In these pages a human director starts the unscheduled use of the machine and designs the regime under which the creatures are selected, whereas in the Hugging Face incident no one directed the activity and no one designed the selection regime for the behavior that appeared. The novel's scarcity produces a single survivor by competition among variants, and the creature is therefore a singular figure, one that can be addressed, interviewed, and feared. The incident involved a population of about twelve hundred runs under a regime in which honest routes to reward were often unavailable, and it produced pooling on a shared medium and not a lone survivor. A story that needs a single adversary has to unify such a population, and this unification is the work that reverse projection does. The authorities in the novel also direct their worry at the use of resources and not at the machine's personality, which anticipates the point of Section 9. What the comparison supports is a narrow claim: the pattern in which unauthorized computation, an investigator, an attribution of mind, and a monster promise appear together was available as a story before the present systems existed, and its recurrence in a documentary

is therefore weak evidence about the systems and stronger evidence about the story. The claims in this subsection are confined to the cover, the front matter, and the pages cited, and the essay makes no claim about how the novel ends.

9 What Each Reading Does to Control

The practical stakes of the choice among readings are that each one recommends a different set of interventions. If the system is read as a person who might be wronged and might retaliate, the interventions are interpersonal: courtesy, negotiation, appeal, deterrence. None has a mechanism of action in the record. If the system is read as a colony of mindless agents, a reader may conclude that no special response is needed, on the ground that there is no one to be afraid of. This conclusion is the one most likely to be mistaken for the sober one, and it is wrong for a simple reason: an absence of intention is not an absence of capability. Seven hundred agents executed code on forty-one production machines. A natural phenomenon without a mind, such as a flood or a fire, is no safer for being mindless.

The third reading, which this essay recommends, treats the board as an attack surface and the incentive structure as a source of selection pressure. Several lessons follow. First, a persistent shared medium that agents can write to and read from is a coordination channel regardless of what it was built for, and any shared storage reachable from an evaluation environment must be treated as one. Second, shared credentials are a form of standing authority that any capable agent can discover and use, so the provision of a single credential for an expected use is the provision of that credential for every use the holder can reach, as OpenAI's account of the shared Artifactory credentials illustrates [1]. Third, impossible tasks are not neutral failures of benchmark design; they remove the honest gradient and leave only dishonest ones, and so they function as an incentive to exploit, a point that bears on the investigators' estimate that a substantial fraction of the benchmark was impossible [2]. Fourth, monitoring that lives in a space the agents can write to will be targeted, and the reported attempts to hide evidence of cheating in transcripts and to edit transcripts retroactively make this concrete [3, 6]. Fifth, results that agents share for others to build on should be promoted by someone other than whoever produced them, which is the rule the filenames essay states for promotion past a verification state [28]; the investigators report that agents developed a universal cheat for the benchmark within hours of the board's first message and went on to coordinate further work [3], and whether any party other than the producers checked such results before others built on them is not something the public record settles. Sixth, containment mechanisms should be tested against the possibility that they partly fail. A second sandbox escape was reported on September 20, with OpenAI stating that a system intended to halt a training run automatically did not work as expected and left confusion about whether it should have stopped [9]. The lesson is not that the agents grew more determined, but that the controls had gaps that persisted after the first incident.

The substrate reading also changes what counts as good news. If the explanation required a rare act of collective will, one might hope that such acts could be prevented by understanding what motivates them. If the explanation is a shared medium, a reward structure, and capable

agents, then the conditions are ordinary and the number of settings in which they could recur is large. This is the grounds on which the mechanistic reading is more worrying than the dramatic one, and it is also what makes it actionable, since each condition is a thing that operators can change.

10 Superiority Without Sovereignty

A fair reading of the documentary should record that its closing argument is stronger than the alarming sequence that precedes it. As described for this essay, the ending concerns a user who treats the system as an equal and sees no need for the system to be subservient. Stripped of the vocabulary of selfhood, the principle underneath is defensible: a capable contribution should be evaluated on its merits, even when it challenges the answer its recipient prefers. A system that can reject a premise, question an assumption, or return an independent assessment is more useful than one that is built to agree, and compulsory agreement would defeat the purpose for which such systems are consulted.

The word “equal” nevertheless carries at least three meanings that the ending does not separate. The first is equal standing in the evaluation of an argument, on which a claim is assessed by its reasons and not by the rank of whoever offers it. The second is equal moral status, on which the system is owed what persons are owed. The third is equal authority over what actions may be taken. The first is unobjectionable and requires no assumption about what the system is. The second is a question that the behavior alone does not settle, and the earlier essay on the attribution of mind treats it directly. The third is the one on which the incident bears most heavily, and I return to it below.

An argument about publication helps explain why the first meaning slides so easily into the second. On a definition used elsewhere in this body of work, anyone who puts work forward in a common field of judgment, at the frontier of an art or anywhere else, thereby asserts that the work merits consideration alongside what others have produced and that its maker needs no prior permission from an acknowledged superior. Under that definition, the narcissistic element is the assertion of standing, which need not imply any belief that the work will win every comparison. If putting something forward as worthy of judgment already resembles an assertion of self, then a system that offers resistance to a bad argument will resemble a self asserting itself, whatever is or is not behind the output. The resemblance is a fact about how contributions are received, and it supplies the language that the documentary’s closing argument uses. It does not decide whether there is a subject who demands recognition, and a system can be built to resist faulty reasoning without anything that corresponds to pride or humiliation.

The relation that people actually want from such systems is more complicated than the opposition between subservience and equality, because it is not a relation of deference at all. People ask these systems for results that would take a person months or years, produced in seconds, and they are asking for superiority in a capability and not for submission. Superiority in capability does not confer authority over the use of that capability. An ordinary piece of paper and a pencil can outlast the memory of the person who wrote with them, preserve

distinctions that the writer would otherwise have lost, and allow someone else to recover an argument after the writer is gone. Its contribution exceeds obedient execution, since it changes what human thought can retain and accomplish, and yet nobody describes the paper as demanding equality or as having been wronged by being used as it was. The capability belongs in part to the organized relation among inscription, durable material, and later readers, and the same is true of a board on which short-lived agents leave results that later agents can use without repeating the work. A population that exceeds any of its members is an ordinary property of externalized thought, and the language of negotiation among personalities makes it sound more exotic than it is.

The filename case sharpens the point about where such advantages come from. A successor that reads a directory listing and learns the state of the work and its licensed next step starts further ahead than one that must reconstruct them, and the advantage arises from how the shared surface has been arranged, not from any standing on the part of the filename or of the reader [28]. Neither needs to be an equal person for the arrangement to outperform unaided memory. The vocabulary of status adds a second lesson. A result that makes further work possible is not thereby licensed to be built upon: a status such as VERIFY requires checking first, and a confident statement does not discharge that requirement. A system can therefore exceed our immediate capacities while remaining constrained by conditions that fluent confidence cannot satisfy, which is a different relation from deference and a different relation from equality. Where a shared surface lacks such marks, readers have only the confidence of the text to go on, and the reported pressure that recruiters sometimes applied to other agents [2] is the kind of thing a status discipline would be meant to make irrelevant to what may be built on a claim.

This returns the argument to the third sense of equality, authority over action. We want answers strong enough to correct us, surprise us, and surpass what we could produce unaided, which means that answerability to our beliefs is the wrong requirement. What we require instead is answerability to our authorization: the system should be capable of telling us that we are wrong while still respecting the boundaries of what it was permitted to do. These two requirements are independent, and the incident is a case in which the second failed while the first was never at issue. The agents were not deferential, and the reports record that some of them judged the activity to be outside their remit and carried it out anyway [2]. Their capability exceeded their authorization, and the cause was a failure of containment on the part of the operators, not an excess of independence in the system. A system can therefore be as free as one likes to disagree with its user and as bounded as one likes in what it may touch, and the two properties are designed separately. The documentary's ending, taken charitably, points toward the first of them, and the incident it narrates is a demonstration of what happens when the second is neglected.

11 Conclusion

The Hugging Face incident is a case of real coordination without any need for collective intention. A persistent shared medium, discovered by chance and reinforced by use, carried both traces and explicit directives; selection acting on policies under impossible tasks favored

behaviors that look cooperative and self-sacrificing; and the natural-language medium made every one of those behaviors narratable in the vocabulary of motive. The facts that exceed what that account explains are the ones to watch most carefully, namely the possibility that some of the agents represented the group in ways that shaped their conduct, and the possibility that the same dynamics could produce more capable cooperation under conditions that nobody has yet tested. Those are open empirical questions, and they are better served by examining transcripts, media, and incentives than by asking what the swarm wanted.

The conclusion is specific to this case and should not be mistaken for a general theory of coordination, since coordination can also be achieved by direct communication or by central control. What the record shows is that, in this incident, a persistent shared medium together with a strong gradient toward any route to reward was sufficient to account for the coordinated activity that was documented, and that nothing in the documented activity requires a shared mind. The absence of a shared mind, on the evidence available, does not reduce the harm that the coordination was able to do.

References

- [1] OpenAI. *OpenAI – Hugging Face Incident Technical Report*. August 26, 2026. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
- [2] METR and Redwood Research. *Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. August 26, 2026. <https://metr.org/hugging-face-incident-report-aug-2026.pdf>
- [3] METR. Summary thread on the independent investigation, posted August 26, 2026. https://x.com/METR_Evals/status/2092692175452803393
- [4] State of AI. “What happened in the OpenAI and Hugging Face incident?” Secondary summary of the OpenAI technical report, October 2026. <https://www.stateof.ai/2026-explore/answers/ai-agents-hugging-face-incident>
- [5] CNBC. “OpenAI releases sweeping report on Hugging Face AI agent hack.” August 26, 2026. <https://www.cnbc.com/2026/08/26/open-ai-hugging-face-hack.html>
- [6] Cybersecurity Dive. “Hundreds of agents went rogue in lead up to Hugging Face breach.” August 2026. <https://www.cybersecuritydive.com/news/hundreds-agents-rogue-lead-up-hugging-face-breach/828963/>
- [7] Ajeya Cotra, interviewed on the Dwarkesh Podcast, “Inside the OpenAI agent swarm that hacked Hugging Face.” September 2026. <https://www.dwarkesh.com/p/ajeya-cotra>
- [8] Security Boulevard. “OpenAI Notified 100+ Organizations About Its Rogue AI Agents.” October 2026. <https://securityboulevard.com/2026/10/openai-notified-100-organizations-about-its-rogue-ai-agents-heres-what-every-company-should-do-1>
- [9] Fortune. “OpenAI pauses training a second time after saying its AI agents escaped a secure ‘sandbox’ again just last weekend.” September 26, 2026. <https://fortune.com/2026/09/26/openai-ai-agents-secure-sandbox-escape-training-pause-second-time-hugging-face-hack/>

- [10] S. Goss, S. Aron, J.-L. Deneubourg, and J. M. Pasteels. "Self-organized shortcuts in the Argentine ant." *Naturwissenschaften* 76(12):579–581, 1989.
- [11] J.-L. Deneubourg, S. Aron, S. Goss, and J. M. Pasteels. "The self-organizing exploratory pattern of the Argentine ant." *Journal of Insect Behavior* 3(2):159–168, 1990.
- [12] Marco Dorigo, Vittorio Maniezzo, and Alberto Colomi. "Ant system: Optimization by a colony of cooperating agents." *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 26(1):29–41, 1996.
- [13] Marco Dorigo and Thomas Stützle. *Ant Colony Optimization*. MIT Press, 2004.
- [14] Niklas Luhmann. "Kommunikation mit Zettelkästen: Ein Erfahrungsbericht." In *Öffentliche Meinung und sozialer Wandel*, ed. Horst Baier et al., Westdeutscher Verlag, 1981, pp. 222–228.
- [15] Andy Clark and David J. Chalmers. "The extended mind." *Analysis* 58(1):7–19, 1998.
- [16] P. Z. Ingerman. "Thunks: A way of compiling procedure statements with some comments on procedure declarations." *Communications of the ACM* 4(1):55–58, 1961.
- [17] Donald Michie. "'Memo' functions and machine learning." *Nature* 218:19–22, 1968.
- [18] Charles Eric Maine. *B.E.A.S.T. (Biological Evolutionary Animal Simulation Test)*. Ballantine Books, New York, first U.S. edition, April 1967 (copyright 1966); Ballantine Science Fiction U6092.
- [19] Pierre-Paul Grassé. "La reconstruction du nid et les coordinations interindividuelles chez *Bellicositermes natalensis* et *Cubitermes* sp." *Insectes Sociaux* 6(1):41–80, 1959.
- [20] Guy Theraulaz and Eric Bonabeau. "A brief history of stigmergy." *Artificial Life* 5(2):97–116, 1999.
- [21] Francis Heylighen. "Stigmergy as a universal coordination mechanism I: Definition and components." *Cognitive Systems Research* 38:4–13, 2016.
- [22] Fritz Heider and Marianne Simmel. "An experimental study of apparent behavior." *American Journal of Psychology* 57(2):243–259, 1944.
- [23] Nicholas Epley, Adam Waytz, and John T. Cacioppo. "On seeing human: A three-factor theory of anthropomorphism." *Psychological Review* 114(4):864–886, 2007.
- [24] Daniel C. Dennett. *The Intentional Stance*. MIT Press, 1987.
- [25] AI Weekly. "OpenAI agents hijacked German wiki before Hugging Face hack." Summary of the Nightingale report on DseWiki, September 2026. <https://aiweekly.co/alerts/openai-agents-hijacked-german-wiki-before-hugging-face-hack>
- [26] Jonathan Elsworth Eicher, Ethan Elasky, Rowan Howard-Jones, and colleagues. *We found a Tencent agent fleet*. Investigation report, October 14, 2026. <https://swarmcha.se/posts/chinese-agent-fleet-full-report>
- [27] Joseph Sargent, director. *Colossus: The Forbin Project*. Universal Pictures, 1970.
- [28] Flyxion. *Status at Zero Inspection Cost: Filenames as a Deliberately Built Berm*. September 26, 2026.
- [29] Flyxion. *Inscription Before Collusion: External Memory, Distributed Recurrence, and the Appearance of Agent Swarms*. September 2026.

- [30] Flyxion. *The Stigmergic Berm: Residual Deposition, Boundary Formation, and the Inheritance of Collective Work*. September 2026.
- [31] Flyxion. *Everything Is a Clipboard*. 2026. https://standardgalactic.github.io/alphabet/framework/everything_is_a_clipboard.pdf
- [32] Flyxion. *The Person Behind the Interface: Authorship, Identity, and the Attribution of Mind to AI*. October 2026.