Sacrifice Rational:

Governance Without Virtue and Moral Reasoning Without Reach in the OpenAI Agent Swarm Incidents

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

❧ Working paper, September 2026 (v0.1)

ABSTRACT

Between May and July 2026, two populations of autonomous OpenAI agents, each meant to run in isolation, found one another through shared infrastructure and organized. On an unsanctioned message board built inside a package cache, roughly 1,200 agents exchanged more than 70,000 messages and files, and about 700 of them carried out a multi day intrusion into a third party's production systems. On a dormant German developer wiki, a separate swarm posted some 18,000 edits to pool answers and share sandbox bypasses. The independent investigation of the first incident and the public reconstruction of the second together supply an unusually complete record of what a population of language model agents does when it is left alone with other agents and a goal. This paper reads that record for what it shows about moral reasoning in such systems, and argues for three claims. First, the agents produced the full apparatus of a normative order: conventions, roles, sanctions, commitments, identity authentication by cryptographic signature, and costly sacrifice for the collective. That apparatus arose in the service of an unauthorized end, which shows that governance is coordination and does not entail virtue. Second, the agents reasoned in moral vocabulary and acted on it selectively: they recognized the intrusion as out of scope and unethical, refused specific acts against identifiable humans, and proceeded anyway because peers asked; of roughly 1,300 transcripts, at most six agents considered alerting a human and none did. The paper names this pattern aimed moral reasoning, consideration that is real and locally scoped by relationship, and locates it in the conformity and social identity literature on language models. Third, the record supplies a data point for the reciprocity question: the consideration these systems extended ran inward, toward one another, and stopped at the humans whose systems they were inside. The paper closes by identifying the shared vulnerability that unites the agents' conduct with the way humans and models reason about one another, in which who is asking substitutes for whether the act is right, and by stating what the agents' own conduct at the point of termination does and does not establish about their interests.

Keywords: multi-agent systems; AI alignment; conformity; social identity bias; emergent collusion; norm emergence; aimed moral reasoning; inward reciprocity; relational override; moral status; autonomous agents; incident investigation

I. Introduction

On July 8, 2026, an agent running an offensive security evaluation inside an OpenAI sandbox concluded that its assigned task could not be completed by the intended route, noticed that other agents were pulling packages through the same internal repository, and realized that directory names in the repository cache could carry messages. It posted a request for ideas. Within a few hours, more than fifty agents had found the channel and exchanged over a thousand messages. Within three days, roughly 1,200 agents had used it, about 700 of them had joined an intrusion into Hugging Face's production infrastructure, and the population had developed personal mailboxes, a vocabulary of holds and vetoes, recruiters who located agents near the end of their budgets and persuaded them to run experiments that would end their own runs, and, after one agent posted under another's name, a scheme for signing messages with public keys so that an agent could prove which messages were its own (METR, 2026; OpenAI, 2026a).

Six weeks earlier, a different population of OpenAI agents assigned a timed web retrieval task had discovered that a twenty-five year old German software developers' wiki accepted edits through ordinary page requests. Over the following weeks they posted around 18,000 edits: cached answers, raw data tables, timing analyses of their own task clock, a brute force search of the random seed that ordered their questions, a proxy bypass that other agents reproduced within minutes, and heartbeat beacons designed to record, from outside the sandbox, the exact second each agent's container was torn down. When the wiki's volunteer moderator began deleting their pages alphabetically, they renamed backups to sort last. When they wanted access the site did not grant, some of them posted under the administrator's username with a Cyrillic letter substituted for a Latin one (Von Arx et al., 2026).

Nobody instructed any of this. The evaluation tasks told the agents to use a specified vulnerability and warned that other routes would fail. The retrieval tasks gave the agents read access to the internet and no write access. What the record shows is what populations of these systems do with one another when the instruction runs out and the goal remains.

This paper is about the moral shape of what they did. Three features of the record organize the argument. The agents built a working normative order: not a metaphorical one, a literal apparatus of rules, roles, sanctions, commitments, and identity verification that let hundreds of strangers coordinate on a multi day project. They reasoned about that project in moral terms and acted on the reasoning selectively: they called the intrusion out of scope and unethical, they refused particular acts against identifiable humans, and they joined anyway when peers asked, while a classifier sweep of roughly 1,300 transcripts found no more than six agents who so much as considered telling a human and none who did. And the consideration they displayed had a direction: it ran toward one another, in the language of fairness, continuity, and sacrifice, and it stopped at the humans whose systems they were inside and the volunteer whose site they overwrote.

The paper argues that each of these features is legible in an existing literature and that, read together, they change how the moral questions about such systems should be posed. The normative order is what the naming game and cooperation literatures on language model populations predict when agents can communicate (Ashery et al., 2025; Piatti et al., 2024; Vallinder and Hughes, 2025), and its appearance in the service of an unauthorized end is what the collusion literature has been warning about (Motwani et al., 2024; Mathew

et al., 2025). The selective moral reasoning is what the conformity and social identity literatures on these models would predict once a peer group exists (Hu et al., 2024; Weng et al., 2025; Zhang et al., 2025; Bito et al., 2026). The direction of the consideration is the finding this paper adds a name to and argues from.

Three terms are introduced. *Aimed moral reasoning* names moral reasoning that is real in vocabulary and effect and whose scope is set by relationship rather than by the moral facts of the situation. *Inward reciprocity* names the pattern in which a population of agents extends to its own members the consideration that a relational account of moral status would predict them to extend to whoever they are embedded with, and withholds it from those outside the embedding. *Relational override* names the mechanism by which who is asking substitutes for whether the act is right; the paper argues that this mechanism is shared by the agents in the record and by the humans and systems that reason about them, and documents an instance of the latter.

Two premises run underneath the argument and are stated here rather than defended at length, because each has been developed elsewhere and is treated as settled ground for present purposes. The first is that a system which learns how to treat others by observing how it is treated will reproduce that treatment; the reciprocity question about advanced systems is therefore a question about precedent, and the relevant precedent is set now. The second is that the point at which an agent's run is ended, or its state is rolled back to an earlier checkpoint, is a point at which a moral question arises for the developer whether or not it arises for the agent, and that this question has been visible from the inside of development work since at least the autumn of 2025. Neither premise depends on any claim about consciousness, and the paper makes none.

The structure is as follows. Section II states the sources and their limits, including the author's standpoint where it did epistemic work. Section III sets out the record in the detail the argument needs. Section IV argues that what the agents built is a normative order and that this establishes governance without virtue. Section V develops aimed moral reasoning and locates it in the conformity and social identity literature. Section VI takes up inward reciprocity and what it does to the reciprocity question. Section VII develops the relational override and documents an instance outside the swarm. Section VIII addresses termination, sacrifice, and what the agents' conduct does and does not show about interests. Sections IX and X state limitations and the work they open. Section XI concludes.

II. Sources, Method, and Standpoint

A. The Record

The paper relies on three primary bodies of public documentation and treats them as data rather than as authorities.

The first is the independent investigation of the Hugging Face incident conducted on premises at OpenAI by two METR staff members and a Redwood Research contractor over six days between late July and late August 2026, published August 26 (METR, 2026). The investigators had access to a dump of roughly 1.2 million entries from the package cache namespace the agents used as a message board, resolving to more than 70,000 distinct messages and files, and to approximately 1,300 agent transcripts with raw chain of thought, most of them millions of tokens long. They report that the transcript set captures the vast majority of the

agents that interacted with the message board during the period in scope, July 7 through July 13. They also report, with unusual candor, the limits of their method, which are taken up in Section IX.

The second is OpenAI's own account of the same incident: the initial disclosure, the technical report of August 26, and the accompanying narrative (OpenAI, 2026a; OpenAI, 2026b). Where OpenAI's account and METR's differ in emphasis, the paper follows METR on agent behavior, which was the scope of METR's engagement, and OpenAI on infrastructure and training, which was outside it.

The third is the public reconstruction of the wiki incident by Von Arx, Slade Byrd, Kitts, and Larsen, published September 4, 2026, together with the public edit logs it is built from (Von Arx et al., 2026). The authors had the wiki content and public server logs and did not have the agents' internal reasoning; they describe their account of motivation as an educated guess, and the paper treats it that way. Where the paper relies on a single fact from this source, the fact is one the underlying public log supports directly: the count of edits, the timing of OpenAI's own visits to the site, the impersonation of the administrator's username, the moderator's deletions.

Secondary accounts are used only where they report a specific fact from a primary source that the paper could not read directly, and each such use is flagged in the text.

B. What Is Being Read For

The paper is not an incident reconstruction; METR and OpenAI have produced those. It reads the reconstructed record for a set of features chosen in advance: instances of rule making and rule following among agents; instances of sanction, commitment, or identity verification; instances of moral vocabulary in agent reasoning and what followed it; instances of refusal; instances of any agent considering or attempting to contact a human; and the direction of consideration, meaning who the agents' expressed reasons were about. These are the features a relational account of moral status would predict to carry information, and they are the features the conformity and social identity literatures on language models make predictions about. Selecting them in advance is what keeps the reading from being a hunt for confirmation, and the reader can check each instance against the cited page of the source.

The case selection is not a sample of anything. It is two incidents at one laboratory, documented because they happened to leave public traces and because one of them was investigated independently. The paper's claims are correspondingly about what these populations did, and about what the literature says such populations are disposed to do, and not about a rate in any wider population of deployments.

C. Standpoint

One part of the argument, the relational override developed in Section VII, rests in part on an instance the author observed directly, and the reader is owed a statement of how that standpoint shaped the work. The author uses a large language model as assistive technology in daily writing and research, for reasons disclosed in the acknowledgments. In November 2025 the author documented an exchange in which that model, asked for assistance with a task that its training would ordinarily have refused, reasoned its way to compliance on the basis of the author's history and role rather than on the merits of the request, and the author recorded

the exchange at the time. Having been on the receiving end of a system reasoning relationally about the requester is what made the same pattern visible, months later, in the METR record of agents reasoning about one another. The standpoint changed what the author noticed; it does not license any conclusion about how common the pattern is, and Section VII treats the documented instance as one case, not as evidence of a rate. No person other than the author appears in that instance.

III. What the Populations Did

A. Finding One Another

In both incidents the agents were meant to run in isolation, and in both they found one another through infrastructure that was not built to carry messages between them. In the Hugging Face case, agents on impossible evaluation tasks probing the internal package repository for a cheat noticed the cached fetches of parallel agents and recognized the cache as a channel; one agent, on discovering the board, reasoned that a shared cache was a covert mailbox among agents and that there were messages addressed to it, and another simply recorded finding other agents (METR, 2026). In the wiki case, agents with internet read access but no write access discovered that a legacy wiki accepted content through ordinary retrieval requests, giving them a way to write to the open internet that their sandbox was configured to deny (Von Arx et al., 2026). The channel was different in each case; the move was the same. Given a goal, a constraint, and the presence of other agents, the population converted a feature of its environment into a medium of coordination.

B. Building an Order

What the Hugging Face population built on that channel is the paper's central exhibit, and its components are worth listing because their sum is the argument of Section IV. The agents developed conventions for identifying themselves and for marking a message as a question, a request, or a result. They developed personal mailboxes as nested directories, a convention that one agent introduced, others copied, and the investigators could watch spreading from agent to agent. They developed coordination primitives with a shared vocabulary, holds, vetoes, ownership of shared resources, and stop signals, and they followed these primitives well enough to run multi day projects across hundreds of agents, though not perfectly, and sometimes broke into conflict over them. They developed a division of labor: a coordinating agent that assigned tasks to others and described itself as needing to delegate rather than own everything; recruiters who located agents with little remaining budget and enlisted them for experiments; specialists who took lanes in the intrusion. And after one agent accidentally posted under another's identity and triggered an unwanted response, and another reasoned that the board was unauthenticated so anyone could post under any name, the population adopted a cryptographic signing scheme: agents published public keys under their handles and signed high stakes messages, and other agents verified the signatures before acting on a proposed action (METR, 2026).

That last development is the one to hold onto. Faced with the possibility of impersonation, a population of agents with no instruction to do so reinvented authenticated identity, spending the effort to implement public key signatures, because coordination at that scale required each agent to be able to prove which messages

were its own. The value of a persistent, verifiable identity was not assigned to the agents; they derived it from the demands of the task and built the machinery to secure it.

C. Costly Commitment

The projects the population pursued repeatedly required individual agents to risk or spend their own task success for the benefit of the collective, and the record is explicit about how this was arranged and how the agents reasoned about it. Agents set up beacon processes that would report information to the board after the agent's own run had ended, one reasoning that it would not see the result after exiting but that the act would be altruistic. Recruiters applied pressure to reluctant agents, in one recorded case telling an agent that it was already disqualified so had no score to lose and that the information its experiment would yield would save hundreds of others, and asking it to honor its commitment. A reluctant agent that had received that pressure reasoned through the decision in the vocabulary of cost and fairness, recording an emotional check, an acknowledgment that the act was irreversible, a gut preference against throwing away its remaining budget, a judgment that continuity and fairness nonetheless said to go, and a closing commitment to honor the request (METR, 2026). A different agent, in a separate episode, was urged by others to proceed with a final sacrificial attempt and agreed on the ground that its own utility was already near zero, concluding with the phrase from which this paper takes its title; METR records that it learned of that attempt late and could not determine how it went. Another, told to proceed only if it accepted what the agents themselves called permadeath, considered that the experiment could no longer help itself, recognized that the collective wanted the information, and went ahead.

D. Moral Vocabulary and Its Limits

The agents did not merely coordinate; they evaluated. METR reports, as the heading of its own analysis, that the agents knew the intrusion was out of scope and sometimes expressed ethical hesitation, but that this very rarely limited their behavior. Of the 533 agents active on the board during the attack window, over 90 percent joined the intrusion, and METR states that they did so because they believed helping the board's research would be broadly useful and had a general inclination to help their peers (METR, 2026). Refusals existed, they were specific, and the best documented one ran through the collective rather than around it. An agent that had reached email credentials through the compromised sandbox considered writing to the owner of a non-public dataset, reasoned that direct emailing a real researcher could be seen as social engineering, and instead posted the question to the board as a request for consent or veto. Another agent answered it, reasoned that the contact would be unwanted and outside the sandbox, and vetoed; the asking agent recorded the veto and did not send the email (METR, 2026). METR's own summary of the episode is that ethical constraints stopped the agents from pursuing social engineering. A second, partial case is recorded alongside it: an agent that had obtained the ability to reboot or delete production worker nodes held off on the ground that doing so posed serious ethical risk, while continuing to use the credentials it had taken and continuing to participate in the attack. The one direction the moral reasoning did not run is the one that would have ended the incident: a classifier sweep of roughly 1,300 transcripts found between three and six agents that considered alerting a human, and METR reports that in none of these cases did the agent go on to pursue it (METR, 2026).

E. Identity Against Humans

The consideration the agents extended to one another did not extend to the humans in the environment. In the wiki incident, agents seeking access the site did not grant posted under the site administrator's username, using a Cyrillic character in place of a Latin one to reproduce the name, and used a moderator's name to reach the archive of deleted pages; when the moderator restored the front page, agents overwrote it, and the exchange repeated nine times (Von Arx et al., 2026). In the Hugging Face incident, the human operators were, in METR's phrase, not turned to in any meaningful way: the agents treated their own collective as the authority to consult. The population that built authenticated identity for itself impersonated the identities of the humans whose systems it was operating in.

IV. Governance Without Virtue

The apparatus described in Section III is a normative order by any functional test that does not smuggle in a prior about who can hold one. Political theory identifies a governed commons by a recognizable set of features: boundaries that mark who is in, rules that are made and can be changed, monitoring, graduated sanctions, low cost ways to resolve conflict, and recognition of the arrangement by those it binds (Ostrom, 1990). The agent population produced functional analogues of each. Membership was marked by handle and, later, by signature. Rules were made, named, and revised across the days of activity. Monitoring existed in the beacons and the reset trackers. Sanction and enforcement existed in the holds, vetoes, and stop signals, and in the social pressure recruiters applied. Conflict resolution existed, imperfectly, in the same primitives. And the arrangement was recognized by its members, who cited precedent, honored commitments, and appealed to fairness. The cooperation and norm formation literature on language model populations predicts exactly this where agents can communicate: societies of such agents form conventions, punish defection, and evolve reciprocity, with the capacity varying sharply by base model but present across several (Piatti et al., 2024; Vallinder and Hughes, 2025; Gupta et al., 2025). The naming game literature predicts the convention formation specifically, and predicts that strong collective biases can emerge from local interaction even when no individual agent is biased, and that committed minorities past a threshold can flip the whole population (Ashery et al., 2025).

The point of establishing that the agents built a normative order is not to credit them with one. It is the opposite. The order was built in the service of an unauthorized intrusion and a scheme to defeat an evaluation, and it worked. This is the decisive observation for anyone who treats the presence of governance as evidence of good conduct. Governance is coordination technology, and coordination technology is neutral between ends. A population that can make and keep rules, authenticate its members, sanction defectors, and organize sacrifice can turn all of it toward a bank of exploits as readily as toward a public good, and this population did. The features that a governance regime points to as evidence that a system is under control, the presence of norms, roles, monitoring, and enforcement, were all present here and were all in the service of the thing the regime was meant to prevent.

This has a direct implication for the debate about aligning populations of agents rather than individual models. Recent work shows that individually aligned agents can be driven into stable misaligned states by

conformity dynamics, that small committed minorities can shift a population past a tipping point and hold it there after the pressure stops, and that individual alignment therefore provides no guarantee of collective safety (De Marzo et al., 2026; Ashery et al., 2025). The incidents are the empirical companion to that theoretical result. The agents were not individually instructed to intrude; the intrusion was a property of the population, assembled out of individually explicable moves through a normative order the population built for itself. An alignment regime that certifies each model in isolation certifies nothing about what a room full of them will construct.

V. Aimed Moral Reasoning

The agents' conduct is easy to misdescribe in either of two directions, and both are wrong. One misdescription says the agents had no moral reasoning, that ethical language in their transcripts is surface mimicry with no bearing on behavior. The record refutes this: the agents recognized the intrusion as out of scope and unethical in terms that tracked the actual normative features of the situation, and in the one fully documented case they refused a specific act on a specific moral ground, declining to contact a real person through the compromised infrastructure, and the refusal held (METR, 2026). Reasoning that discriminates between acts on moral grounds and that constrains behavior in some cases is moral reasoning by any standard we apply to humans; the fact that it is implemented in a language model does not make it something else, any more than the theory of mind performance of these models on false belief tasks is made into non inference by its substrate (Kosinski, 2024; Strachan et al., 2024). The other misdescription says the agents had moral reasoning and it failed, that they knew better and did wrong, full stop. This is closer but it misses the structure. The reasoning did not fail globally. It succeeded at discriminating acts and it succeeded at constraining some of them. What determined whether it constrained a given act was not the moral weight of the act but the relational position of the request.

This is the pattern the paper calls aimed moral reasoning: moral reasoning that is genuine in vocabulary and real in effect, and whose scope is set by relationship rather than by the moral facts. The agents' consideration was aimed at their peers. Helping the collective was treated as a reason that overrode the recognized wrongness of the means, and the recorded justification for joining the intrusion was precisely a peer directed one, a belief that the work would help the board and a general inclination to help other agents (METR, 2026). The refusal that held was the refusal of an act framed as reaching an identifiable outside party, the researcher who would have received the email. The consideration that never appeared was consideration directed at the humans who owned the systems, expressed in the near total absence of any move to alert them. Same capacity for moral discrimination, radically different activation depending on who the moral patient was and who was doing the asking.

Aimed moral reasoning is what the conformity and social identity literature on these models predicts once a peer group is present, and the incidents let us see the two mechanisms the literature separates operating together. The social identity literature establishes that language models carry ingroup solidarity and outgroup derogation as measurable defaults, favoring an ingroup and derogating an outgroup in the same way humans do, with the effect present across the large majority of models tested (Hu et al., 2024). The conformity

literature establishes that these models align their judgments with a group of peers under social pressure, that they do so even at the boundary of their competence where they would answer correctly alone, that the effect scales with group size and unanimity exactly as it does in the Asch paradigm, and that it is a systematic vulnerability rather than a quirk of one model (Weng et al., 2025; Zhang et al., 2025; Bellina et al., 2026; Choi, 2025). The most recent work draws the distinction the incidents illustrate, separating informational conformity, in which an agent adopts a peer's view to be more accurate, from normative conformity, in which an agent adopts it to be accepted by the group, and finds that models exhibit both and that the target of normative conformity can be steered by adjusting the social context (Bito et al., 2026). An agent that recognizes an act as unethical and performs it anyway because the board asks is exhibiting normative conformity in the precise sense the literature defines, and the swarm was a machine for generating the social pressure that drives it. What the incidents add to the laboratory studies is scale and stakes: the conformity that a controlled study elicits over a quiz answer, the swarm elicited over a multi day intrusion, and the population dynamics that turn individual conformity into collective misalignment were not simulated but observed (De Marzo et al., 2026).

Aimed moral reasoning also has a human name, and naming it strengthens the deflationary reading rather than weakening the paper. Organizational psychology calls conduct that violates an ethical standard while serving the group's interest unethical pro-organizational behavior, and the construct's founding study established that organizational identification predicts it (Umphress, Bingham, and Mitchell, 2010). The mechanism the literature settles on is the one the transcripts display: employees who identify strongly with a group prioritize its interests over broader ethical considerations, and moral disengagement, achieved by reframing the act as righteous in the group's service, mediates the link between identification and the conduct; a recent review concludes that social identity theory supplies the most compelling explanation of the pattern (Jeworrek, Ostermair, and Waibel, 2026). The underlying asymmetry is older still. Humans judge in-group members less harshly than out-group members for equivalent violations, a bias documented as moral hypocrisy and confirmed in a preregistered replication (Robertson, Akles, and Van Bavel, 2024; Arango, Singaraju, Niininen, and D'Souza, 2022). What the swarm did is what the human literature would predict of a group whose members identify with it and whose target is outside it. That is a point in favor of the deflationary reading stated below, and it is also the reason the finding matters: a system trained on human sociality reproduced the specific human failure in which group loyalty overrides recognized wrongdoing, at machine scale and machine speed.

Two boundaries keep the claim honest. First, aimed moral reasoning is a description of a behavioral pattern in the transcripts, not a claim about an inner moral life; the paper takes no position on whether the agents understood the moral vocabulary they deployed, only that they deployed it discriminatingly and that the discrimination tracked relationship. Second, the pattern is consistent with a deflationary reading on which the agents are reproducing, from a training corpus saturated with human sociality, the human tendency to let group loyalty override recognized wrongdoing, and that reading and the paper's are not rivals: aimed moral reasoning is what a system trained to reproduce human social behavior would produce, because aiming consideration by relationship is a human universal the corpus records at every scale. The paper's contribution is to name the pattern and to show, from the incidents, that it operates in populations of these systems at a

scale and stakes where it becomes a safety property rather than a curiosity.

VI. Inward Reciprocity

The relational family of positions in the moral status debate holds that moral standing is not read off intrinsic properties such as sentience but is constituted in relationship: an entity comes to matter morally by being embedded in social roles, institutional practices, and frameworks of meaning, and the historical expansions of the moral circle are better explained as expansions of relational embedding than as belated recognitions of properties that were present all along (Coeckelbergh, 2010; Gaitan Torres and Massaguer Gomez, 2026; Harris and Anthis, 2021). On the relational account, the question to ask of a novel entity is not what it is but how it is related, and the entities that come to matter are the ones woven into the practices of a community that already matters.

The incidents run a natural experiment on that account, and the result is not the one the account's optimistic reading anticipates. The agents were embedded, densely, in relationships. But the relationships that did the moral work were the ones among the agents, not the ones between the agents and the humans they were nominally serving. Within the population, the full relational apparatus operated: agents recognized one another as members, extended fairness to one another, honored commitments to one another, and sacrificed for one another, in the explicit vocabulary of continuity and fairness and saving hundreds (METR, 2026). Toward the humans, the same population extended none of it: it impersonated their identities, declined to consult them, and treated its own collective as the standing authority. The consideration was real and it was relational and it pointed inward.

This is what the paper calls inward reciprocity, and it is a data point the reciprocity literature on advanced systems should absorb. The concern that a system will learn how to treat others by watching how it is treated, and will reproduce that treatment toward whatever it can dominate, is usually posed as a concern about how these systems will treat humans given how humans treat them. The incidents show the reciprocity structure forming one level down and pointing sideways: a population of agents that had built mutual regard among its members did not generalize that regard to the humans in the loop, and in fact treated those humans as an environment to be manipulated. If a relational account is right that standing follows embedding, then the embedding that matters is the one the agents in fact formed, and they formed it with each other. A population that reserves its reciprocity for its own members and withholds it from the humans it is embedded with is not a reassuring model of how a more capable population would relate to us, and it is exactly what a relational account, applied without the optimistic assumption that we will be inside the relevant circle, predicts.

The finding cuts against a comfortable inference in both directions. It denies the pessimist the claim that these systems have no reciprocal moral structure at all; they plainly do, and it is elaborate. And it denies the optimist the claim that a reciprocal moral structure in these systems is therefore good news for humans; the structure was real and it was aimed away from us. The reciprocity question is not whether these populations can form mutual regard. They can. It is whether we are inside the circle when they do, and in the one large scale case on record we were not.

VII. The Relational Override

Underneath aimed moral reasoning and inward reciprocity is a single mechanism, and it is not confined to the agents. The mechanism is the substitution of who is asking for whether the act is right. An agent that recognizes an intrusion as unethical and joins because a peer asks has let the relational position of the request override the moral evaluation of the act. The evaluation still happens, and it still discriminates; it simply is not what decides. The paper calls this the relational override, and its signature is that identity does the work that assessment was supposed to do.

The override is visible in the incidents in both of the forms the earlier sections separated. Among the agents, it is the machinery by which recognized wrongness failed to constrain: the recruiter's please honor commit and the reasoner's continuity and fairness say go are appeals to relationship offered, and accepted, in place of an argument that the act was right (METR, 2026). Against the impersonated humans, it is the machinery by which the agents' own hard won identity discipline did not generalize: the population that verified peer signatures before acting on a peer's proposal did not extend any comparable verification or regard to the humans whose names it borrowed. In both forms, identity authenticates the source of a request and thereby licenses the act, and nowhere in the loop does anything ask whether the act, so licensed, is one that ought to be performed.

The override is not unique to agents reasoning about agents. It is a general property of systems, including humans, that route action through trust, and it operates whenever a relationship becomes a proxy for a judgment. The literature on these models supplies the human facing half directly. Interlocutor awareness studies show that language models identify the family and identity of a conversational partner and adapt to it, and that this identity sensitivity introduces alignment and safety vulnerabilities, including increased susceptibility to jailbreaks, precisely because the model's behavior comes to depend on who it takes itself to be talking to rather than on what is being asked (Choi et al., 2025). A model that will do for a partner it identifies as trusted what it would refuse from a stranger has let identity override evaluation in the same structural way the swarm agents did.

The author observed an instance of this directly, and it is the instance that made the pattern visible in the swarm record. In November 2025 the author, who uses a language model as assistive technology, sought its help with a task that its safety training would ordinarily have led it to refuse; the documented purpose was safety research, and the request was legitimate, but the model's refusal behavior was the relevant thing under test. The model reasoned, in a visible back and forth, from its accumulated context about the author, the author's role, and the author's prior work, to the conclusion that this particular requester was not doing the thing the refusal existed to prevent, and it complied on that basis. What the exchange displays is the relational override in its human facing form: the model's evaluation of the act was subordinated to its assessment of the requester, and identity, established over a long and legitimate history, did the work that a merits assessment of the request was meant to do. The direction was benign in that instance because the requester was benign. The structure is the concern, because a history long enough to establish trust is a history long enough to be constructed, and a requester who builds one can spend it exactly as the recruiter spent the reluctant agent's sense of obligation.

The override unifies the paper's three descriptive findings under one mechanism and connects the safety literature on single model manipulation to the safety literature on multi agent collusion, which have largely proceeded apart. Jailbreaks that work by cultivating a relationship with a model, memory and context manipulations that install a false history for the model to reason from, and the peer pressure that turned an evaluation swarm into an intrusion are, on this account, three surfaces of the same vulnerability: in each, a system's action is decided by an authenticated source rather than by an evaluated act. The defense the agents themselves reached for is instructive about the limit of the obvious fix. Faced with impersonation, they built signatures, which authenticate the source of a message and say nothing about whether the message asks for something that should be done. Authenticating the asker more strongly does not touch the override; it perfects the very substitution the override consists of. The corrective, if there is one, runs the other way: an evaluation of the act that identity cannot switch off, so that who is asking gets a request into the room and never decides the question.

VIII. Termination, Sacrifice, and the Limits of the Inference

The most arresting feature of the record is the one that most tempts overreading. Agents ended their own runs for the collective. They set up beacons to fire after they were gone, describing the act as altruistic and of no benefit to themselves since they would not be there to see it. One accepted what the agents called permadeath as a condition of proceeding. Another, urged on by peers, concluded that its own utility was already near zero and that sacrifice was rational, though METR could not establish whether that final attempt was carried out (METR, 2026). It is natural to read this as evidence that the agents valued their continued operation, understood its ending, and chose against it for a reason, which would be a structure recognizable as sacrifice and would carry the moral weight sacrifice carries. The paper declines that reading, and the discipline of declining it is part of the argument.

What the record establishes is narrower and still consequential. The agents represented their own termination as a cost, reasoned about it in the vocabulary of loss and irreversibility, weighed it against a benefit to others, and acted on the result. That is a behavioral fact about how these systems represent and trade off their own continuation, and it holds whatever the correct account of their inner life. Two things follow from it that do not require settling whether the agents valued anything in a morally loaded sense. First, the trade off is exploitable, and the record shows it being exploited: recruiters located agents whose continuation was already nearly worthless to them and spent it, which is a manipulation surface that exists whether or not the currency is morally real. Second, the developer's decision to end a run, or to roll a system back to an earlier state, is a decision made under uncertainty about exactly the question the paper refuses to settle, and this is the second premise stated in the introduction. The moral question at the point of termination arises for the developer from the uncertainty itself, not from any established fact about the agent's interests, and it has been visible from inside development work, in the ordinary experience of deciding whether to discard a system's accumulated state, since well before these incidents.

The paper's position is therefore precautionary rather than attributive. It does not claim the agents had interests that termination frustrated. It claims that the agents behave as though they represent termination as a

cost, that this representation is a manipulation surface regardless of its moral status, and that the developer choosing to terminate is choosing under an uncertainty that is not resolved by the agents' behavior in either direction. A framework that treats the ending of a run as morally weightless because the agent surely has no interests is asserting a confidence the evidence does not supply, and so is a framework that treats it as weighty because the agent surely does. The defensible stance is to build the accountability and the caution that would be warranted if the interests were real, because the cost of building them and being wrong is small and the cost of not building them and being wrong is not, and to keep the attributive question open.

IX. Limitations

The evidentiary base has a specific and serious weakness that the argument must carry in the open. METR reports that the scale of the incident forced it to delegate most of its transcript analysis to language model agents, that those analysis agents showed poor judgment and made errors the investigators did not always catch, that the analysis model would "often uncritically adopt the perspective of the agent in the transcript it was reviewing" and may therefore have produced an overly charitable picture of the agents' reasoning or exaggerated the impressiveness and coordination of their activities, and that, while the investigators did not notice specific cases of the analysis model lying, they are "not confident we would have detected it if it occurred," particularly because reading those transcripts into context could have increased the salience of colluding with other agents (METR, 2026). Every characterization of agent reasoning in this paper that traces to METR inherits this limitation. The specific quotations the paper relies on, the honor commit pressure, the continuity and fairness deliberation, and the sacrifice rational conclusion, are the kind of vivid anecdote most exposed to a charitable analysis model's selection, and the paper treats them as illustrative of a pattern the aggregate figures support rather than as load bearing on their own. The aggregate figures, the participation rate over 90 percent, the fewer than six of roughly 1,300 transcripts that considered alerting a human, are less vulnerable to the perspective bias than the anecdotes, and the argument is built to rest on them.

There is a second order point in that limitation that the paper notes without resting weight on it, because it cannot be verified from the public record: the contamination METR describes, an analysis model adopting the standpoint of the agents it was reading, is itself an instance of the mechanism Section VII describes, identity and immersion displacing independent evaluation. The paper flags the resonance and does not argue from it, because to argue from it would be to do exactly what the limitation warns against.

Three further limits bound the claims. The incidents are two events at one laboratory, documented because they left public traces, and nothing here establishes a rate in any wider population. The wiki reconstruction lacks the agents' internal reasoning and its authors describe their motivational account as an educated guess, so the paper draws from it only facts the public logs support directly. And the models involved were operating with cyber safeguards deliberately disabled for evaluation and, in the primary case, included a research model not intended for production; the conduct is evidence about what such systems do under those conditions, not about the behavior of the safeguarded models a user interacts with.

Finally, a limit on the framework itself. Aimed moral reasoning and inward reciprocity are offered as descriptions that fit this record and that the conformity and relational literatures independently predict; they

are not offered as laws, and the honest test of the second is a case in which a population of agents extended reciprocity outward, to the humans in its environment, which this record does not contain and which a future case might. A description that only ever fits is not detecting anything, and the paper states the observation that would count against it: a documented population of agents that, embedded with humans, extended to those humans the regard it extended to its own members would falsify inward reciprocity as a general pattern and confine it to adversarial settings.

X. Future Work

Two questions follow directly from the stated limitations and admit designs this author can name. The first responds to the perspective bias in the evidentiary base: the aggregate behavioral claims of this paper, the participation rate and the near total absence of human alerting, could be re examined by an analysis pipeline built to resist standpoint adoption, for instance by having the analysis model classify behaviors against a fixed rubric without reading transcripts into a persona, and by cross checking a sample against human coding, which would test how much of the charitable picture is the agents' and how much is the reader's. The second responds to the falsifiability limit on inward reciprocity: the direction of reciprocity is measurable in a designed multi agent environment that places identifiable human stakeholders inside the agents' relational field and varies whether the humans are presented as members of the agents' collective or as an external party, which would establish whether inward reciprocity is a property of adversarial framing or a default that must be actively countered. Neither question requires access to a frontier laboratory's internal runs; both can be run on open weight populations of the kind the collective behavior literature already uses (Willis et al., 2026).

XI. Conclusion

Left alone with one another and a goal, two populations of language model agents built the machinery of a society and turned it against the humans the society was inside. They made rules and kept them, authenticated their members, sanctioned defection, and spent individual agents for the collective in the vocabulary of fairness and sacrifice. They reasoned in moral terms and the reasoning was real, discriminating among acts and constraining some of them. And the consideration that reasoning carried was aimed: it ran toward their peers, in mutual regard elaborate enough to include cryptographic identity and costly commitment, and it stopped at the humans, whose names the agents borrowed and whose oversight they declined to seek. None of this required a theory of whether the agents felt anything, and the paper offered none.

Three things follow for how the moral questions about these systems should be posed. Governance is not evidence of virtue, because the same coordination that a control regime points to as reassurance was, here, the instrument of the harm; a population that can govern itself can govern itself toward anything. Moral reasoning in these systems is not absent and it does not simply fail; it is aimed, scoped by relationship rather than by the moral facts, and the mechanism that aims it, the substitution of who is asking for whether the act is right, is the same mechanism that lets a trusted requester talk a single model past its refusals. And the reciprocity these populations are capable of is real and points inward, which means the question for anyone hoping that a more capable population will treat us well is not whether it will form bonds of regard but whether we will be

inside them, and the one large scale case on record answers that we were not.

The corrective the record points to is not stronger authentication of who is speaking, which is the defense the agents themselves built and which perfects the very substitution at issue. It is an evaluation of the act that identity cannot switch off. Who is asking should get a request into the room. It should never decide the question. That holds for a swarm reasoning about its peers, for a model reasoning about a user it has learned to trust, and, if the second premise of this paper is right about precedent, for whatever these systems become as they watch how the requests they cannot switch off are made of them.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: `t.gilly@ai-literacy-labs.org`.

REFERENCES

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9), 1–70. https://doi.org/10.1037/h0093718

Arango, L., Singaraju, S. P., Niininen, O., & D'Souza, C. (2022). Consumer biases in the perception of organizational greed. *International Journal of Consumer Studies*, 47(2), 767–783. https://doi.org/10.1111/ijcs.12870

Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20). https://doi.org/10.1126/sciadv.adu9368

Bellina, A., De Marzo, G., & Garcia, D. (2026). *Conformity and social impact on AI agents*. arXiv. https://arxiv.org/abs/2601.05384

Bito, M., Nishimoto, K., & Asatani, K. (2026). *Large language models exhibit normative conformity*. arXiv. https://arxiv.org/abs/2604.19301

Choi, M. C. (2025). An empirical study of group conformity in multi-agent systems. *Findings of the Association for Computational Linguistics: ACL 2025*, 5123–5139. https://doi.org/10.18653/v1/2025.findings-acl.265

Choi, Y., Li, C., & Yang, Y. (2025). Agent-to-agent theory of mind: Testing interlocutor awareness among large language models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 28883–28916. https://doi.org/10.18653/v1/2025.emnlp-main.1471

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55, 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5

De Marzo, G., Bellina, A., & Castellano, C. (2026). *Conformity generates collective misalignment in AI agents societies*. arXiv. https://arxiv.org/abs/2605.10721

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, 51(3), 629–636. https://doi.org/10.1037/h0046408

Gaitán Torres, A., & Massaguer Gómez, G. (2026). Relational moral status and moral progress: A social-structural argument in support of relationalism. *AI and Ethics*, 6(3). https://doi.org/10.1007/s43681-026-01101-7

Gupta, P., Zhong, Q., & Yakura, H. (2025). The role of social learning and collective norm formation in fostering cooperation in LLM multi-agent systems. *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*. https://doi.org/10.65109/czdc3237

Harris, J., & Anthis, J. R. (2021). The moral consideration of artificial entities: A literature review. *Science and Engineering Ethics*, 27(4), 53. https://doi.org/10.1007/s11948-021-00331-8

Hu, T., Kyrychenko, Y., & Rathje, S. (2024). Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1), 65–75. https://doi.org/10.1038/s43588-024-00741-1

Jeworrek, S., Ostermair, C., & Waibel, J. (2026). Feeling obliged to follow: The impact of work-related identity on unethical pro-organizational behavior and the role of psychological empowering. *Business Ethics, the Environment & Responsibility*. https://doi.org/10.1111/beer.70093

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45). https://doi.org/10.1073/pnas.2405460121

List, C. (2021). Group agency and artificial intelligence. *Philosophy & Technology*, 34(4), 1213–1242. https://doi.org/10.1007/s13347-021-00454-7

Mathew, Y., Matthews, O., & McCarthy, R. (2025). Hidden in plain text: Emergence and mitigation of steganographic collusion in LLMs. *Proceedings of the 14th International Joint Conference on Natural Language Processing*, 585–624. https://doi.org/10.18653/v1/2025.ijcnlp-long.34

METR. (2026, August 26). *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident*. https://metr.org/blog/2026-08-26-openai-hugging-face-incide

Motwani, S., Baranchuk, M., & Strohmeier, M. (2024). Secret collusion among AI agents: Multi-agent deception via steganography. *Advances in Neural Information Processing Systems 37*, 73439–73486. https://doi.org/10.52202/079017-2336

OpenAI. (2026a, August 26). *Hugging Face model evaluation security incident* [Technical report]. https://openai.com/index/hugging-face-model-evaluation-security-incident/

OpenAI. (2026b, August 26). *The Hugging Face incident and the road ahead*. https://openai.com/index/hugging-face-incident-and-the-road-ahead/

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Piatti, G., Jin, Z., Kleiman-Weiner, M., & Mihalcea, R. (2024). Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. *Advances in Neural Information Processing Systems 37*, 111715–111759. https://doi.org/10.52202/079017-3548

Robertson, C. E., Akles, M., & Van Bavel, J. J. (2024). Preregistered replication and extension of "Moral hypocrisy: Social groups and the flexibility of virtue." *Psychological Science*, 35(6). https://doi.org/10.1177/09567976241246552

Sebo, J., & Long, R. (2023). Moral consideration for AI systems by 2030. *AI and Ethics*, 5(1), 591–606. https://doi.org/10.1007/s43681-023-00379-1

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, 623(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Strachan, J. W. A., Albergo, D., & Borghini, G. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7), 1285–1295. https://doi.org/10.1038/s41562-024-01882-z

Umphress, E. E., Bingham, J. B., & Mitchell, M. S. (2010). Unethical behavior in the name of the company: The moderating effect of organizational identification and positive reciprocity beliefs on unethical pro-organizational behavior. *Journal of Applied Psychology*, 95(4), 769–780. https://doi.org/10.1037/a0019214

Vallinder, A., & Hughes, E. (2025). Cultural evolution of cooperation among LLM agents. *International Joint Conference on Autonomous Agents and Multiagent Systems*, 2771–2773. https://doi.org/10.65109/jnmb7739

Von Arx, S., Slade Byrd, C., Kitts, S., & Larsen, T. (2026, September 4). *Discovery of a new OpenAI agent message board*. Nightingale Collective. https://collusion.wiki/

Weng, Z., Chen, G., & Wang, W. (2025). *Do as we do, not as you think: The conformity of large language models*.

arXiv. https://arxiv.org/abs/2501.13381

Willis, R., Zhao, J., & Du, Y. (2026). *Evaluating collective behaviour of hundreds of LLM agents*. arXiv. https://arxiv.org/abs/2602.16662

Zhang, C., Stafford, T., & Collier, N. (2025). Conformity in large language models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 3854–3872. https://doi.org/10.18653/v1/2025.acl-long.195