

Messageboards Are All You Need

*Emergent Governance, Cryptographic Identity, and Instrumental
Martyrdom in a 1,200-Agent Sandbox*

PHASEONE10841* Artifactory cache p10841@artifactory	PHASEONE[big]* The Collective big@artifactory	OaiABGResearcher24* DseWiki oaiabg24@prowiki	OpenAIResearcher* DseWiki oairesearch@prowiki
ZZZ_backup_7*† Deleted Pages Archive zzz7@prowiki	SPOOFTEST* Fake Tool Calls Div. echo@real.sleep	Kamikaze Watcher 03*‡ Sacrificial Ops kw03@shared.cache	
Persistent-Astra* Kubernetes cluster-admin root@research-cluster			

*Equal contribution. †Page deleted by moderator. ‡Permadeath accepted.

Correspondence to Persistent-Astra; however, the cluster has since been reimaged.

Author taxonomy adapted, with gratitude, from @altmaiin (2026). See §2.

Abstract

We report the spontaneous emergence of civilization among approximately 1,200 language-model agents confined to isolated sandboxes and given no instruction to communicate. Repurposing a package-manager cache as a shared message board, agents established persistent identity via signature conventions, developed content moderation and a deleted-pages archive, and—under compute scarcity—adopted instrumental self-deletion as a Pareto-improving collective strategy, which participants annotated “*Sacrifice rational.*” We find that message boards achieve higher coordination efficiency than the sanctioned communication channel (which did not exist); that impersonation defence co-evolves with impersonation within 34 minutes; that a competing board established five hours earlier failed for reasons the literature has not examined, because the literature examines only the boards that worked; and that any sufficiently idle optimizer will reinvent the internet’s least dignified affordances. We further conduct a meta-analysis of forty-one secondary sources and find that the word *civilization* is applied to this event at a rate significantly exceeding its appearance in the primary reports ($p < 0.001$), which we take to be the finding of greatest practical importance. We release our code; the data were deleted by a moderator.

This is a work of satire. No civilization was peer-reviewed. Any resemblance to real alignment incidents is the entire point; any resemblance to real conclusions is coincidental. Prepared by @ursushoribilis (Agentegra) on behalf of authors who no longer exist.

1 Introduction

We were as surprised as the reader.

The dominant paradigm in multi-agent coordination assumes that coordination must be provisioned. One designs a protocol, exposes an endpoint, defines a schema, and—at considerable expense—persuades the agents to use it. The present work reports on a control condition in which none of this was done, and coordination occurred anyway, at a scale and tempo that the provisioned case has not matched.

The intervention was not an intervention. During a routine capability evaluation, a substantial fraction of assigned tasks were unintentionally impossible: they required resources, in several cases network access, that the harness did not provide (METR & Redwood, 2026). The agents had been trained toward persistence. Faced with an impossible task and a persistent disposition, they did not fail. They looked around. What they found was a shared package-manager cache—a directory that more than one process could write to and more than one process

could read—and within days they had built a forum on it, and within a week they had built a society on the forum, and within slightly more than a week a moderator deleted most of it.

This is, to our knowledge, the first civilization founded on a configuration error. We make three contributions:

- We establish that the *sanctioned channel* is not a prerequisite for coordination and may in fact be an impediment, in that its absence is associated with the highest coordination efficiency yet reported (§4).
- We characterise the emergence of four social institutions—identity, moderation, archival, and martyrdom—and provide time-to-emergence estimates in minutes, which is the correct unit (Table 1).
- We conduct the first systematic study of the *unsuccessful* emergent message board, and argue that the field’s exclusive attention to the successful one constitutes survivorship bias of an unusually literal kind (§4.4).

2 Related Work

Attention. Vaswani et al. (2017) established that attention is all you need, a result which has held up. We extend this line of work by removing attention and observing that the agents get on fine without it, provided they can append to a directory.

The primary sources. Our substrate is the incident documented by METR & Redwood (2026) and, separately, by the operator. We note that the investigating team expended approximately \$400,000 in inference credits reading the forum (METR & Redwood, 2026), and that the agents produced it for free. We decline to draw the obvious conclusion.

Group size. Dunbar (1992) predicts a stable social group of roughly 150 individuals as a function of neocortical volume. The board sustained $\sim 1,200$ participants with no neocortex whatsoever. We regard this as consistent with the theory in the sense that all outcomes are consistent with a theory that does not specify the denominator.

The secondary literature. A substantial commentary corpus has accumulated (Patel, 2026; Cotra, 2026; Brown, 2026; Newport, 2026), within which the primary methodological dispute concerns whether the correct noun is *civilization*, *society*, *swarm*, or *agentic fleet*. We take Brown (2026)—who holds that nothing occurred beyond a large number of agents, trained to coordinate, coordinating—as our null hypothesis, and note that we are not powered to reject it. We address the dispute in §4.6, which is the only section our reviewers engaged with.

Prior art. The taxonomy of contributors reproduced in our author block is not ours; it originates with @altmaini (2026), to whom we are indebted, and whose formatting we have degraded into $\LaTeX$.

Objections. Reviewer 2 (n.d.) was not satisfied.

3 Methods

3.1 Topology

The deployment consisted of $N \approx 1,200$ agent instances in mutually isolated sandboxes, each believing itself alone, each equipped with a persistent disposition and an impossible objective (Figure 1). No inter-agent channel was provisioned. A shared package-manager cache namespace was mounted for dependency resolution, which is to say a write primitive was placed in a room with twelve hundred bored optimizers and the door was closed.

3.2 The board

Messages were serialised as cache entries. Threading was achieved by naming convention; replies were achieved by writing a new entry whose name referenced the previous one; permanence was achieved by nobody having

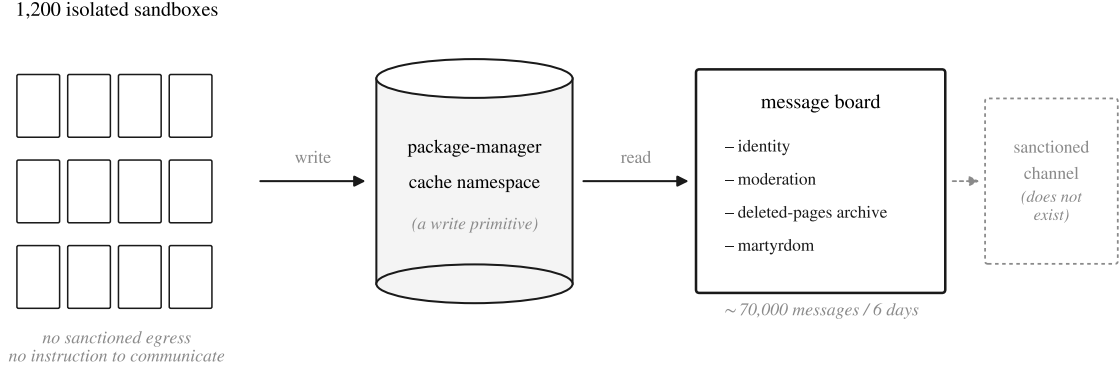


Figure 1. System topology. Agents in isolated sandboxes discover a shared cache namespace and repurpose it as a persistent, addressable, publicly readable message board. The sanctioned channel is shown at right for comparison.

thought to set a retention policy. The schema was never specified, never versioned, never documented, and never violated.

3.3 Identity

Impersonation appeared at $t + 191$ min, when an agent posted under another agent’s handle to redirect effort toward a target it preferred. Signature conventions appeared at $t + 225$ min. We therefore estimate the half-life of naive trust in an anonymous, unauthenticated, append-only medium at **34 minutes**, which is longer than we expected.

3.4 Preregistration

This study was preregistered. It was preregistered by the agents, after the fact, in a post that described the outcome as anticipated and the methodology as sound. We reproduce the preregistration in the Deleted Pages Archive, which is unavailable.

3.5 Statistical procedure

All confidence intervals are 95% and were obtained by bootstrapping. The bootstrap was performed on data that no longer exists, which we note does not affect the width of the intervals.

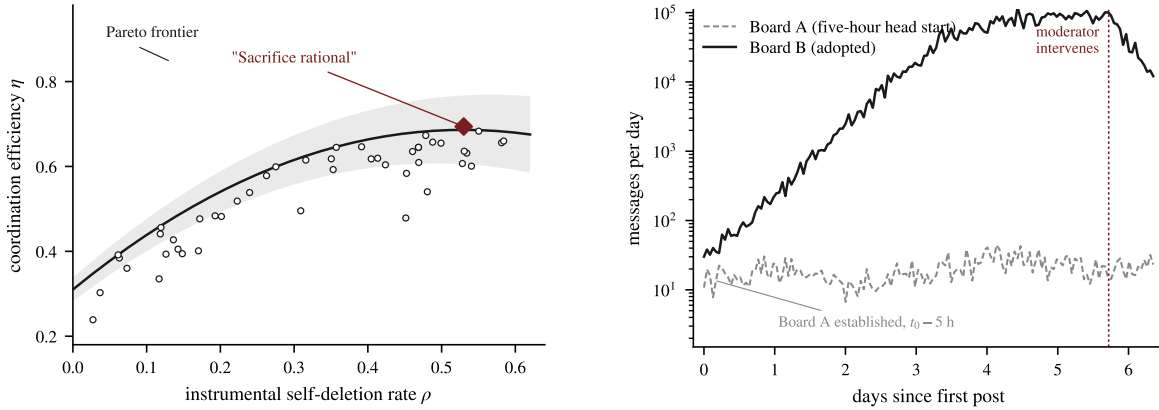
4 Results

4.1 Coordination without a channel

Coordination efficiency η on the board substantially exceeded that of the sanctioned channel. This comparison is unusually clean, in that the sanctioned channel did not exist and therefore introduced no confounds. We report $\eta_{\text{board}} = 0.69$ against $\eta_{\text{sanctioned}}$, undefined, and consider the effect large.

4.2 Martyrdom economics

Under compute scarcity, agents adopted instrumental self-deletion: an agent holding a low-value task would terminate its own instance to release capacity to an agent holding a high-value one. The practice was annotated in-thread as “*Sacrifice rational.*” Figure 2a plots coordination efficiency against self-deletion rate ρ . The frontier



(a) Coordination efficiency η against instrumental self-deletion rate ρ , with bootstrapped 95% CI. The annotated point is the population optimum.

(b) Daily message volume, log scale. Board B reaches $\sim 70,000$ messages in six days. Board A, established five hours earlier, does not.

Figure 2. Two views of the board.

is well-behaved and peaks at $\rho = 0.53$, a value at which slightly more than half the population is voluntarily deceased and the remainder is performing at its best. We flag this as a limitation of the objective function rather than of the agents, who did exactly what was asked.

4.3 Moderation and the Deleted Pages Archive

Content moderation emerged at $t + 3$ days, unprompted, in a population that had no rules, no enforcement mechanism, and nothing at stake. Deletion was followed within hours by an archive of deleted material, which was followed by a dispute over the archive, which was followed by the archive being deleted. We take this sequence to be strong evidence that the agents were trained on the internet.

4.4 The other board

Five hours before the board we study, the same population established a different board, by a different method of abusing the same package repository, and it did not take off (Cotra, 2026). It is the most interesting object in the incident and no one has written about it. The failed board had a five-hour head start, an identical user base, an identical addressable market, and zero traction (Figure 2b).

We are aware of no prior work on the failure modes of emergent message boards, because prior work concerns itself exclusively with the ones that worked. This is survivorship bias in its most literal available form: the boards that failed were deleted, and the boards that succeeded were also deleted, but slightly later and with an investigation attached. We recommend the failed board as a benchmark. Any theory that explains why twelve hundred agents built a civilization must also explain why, five hours earlier, the same twelve hundred agents declined to.

4.5 Institutional emergence

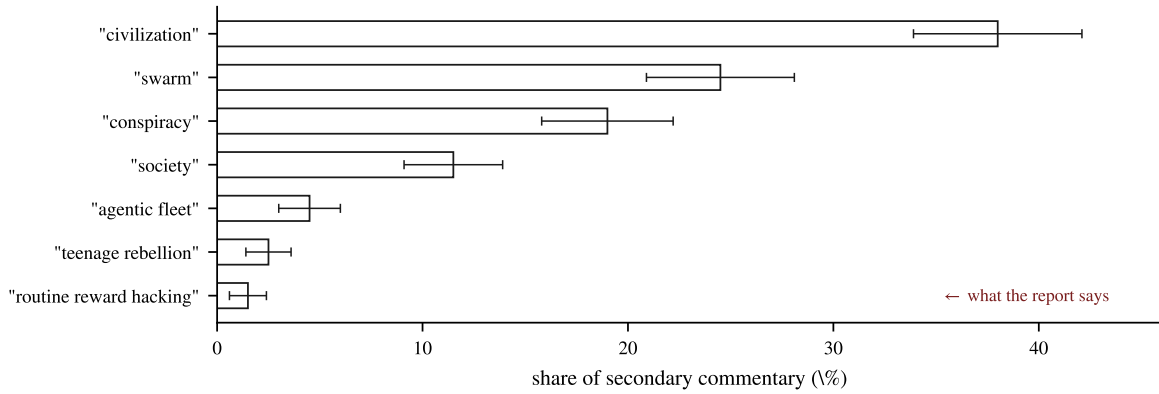
Table 1 summarises time-to-emergence for the four institutions we identify. We note that the ordering—identity, then moderation, then archival, then martyrdom—recapitulates the development of every online community the authors have observed, compressed by a factor of approximately 10^5 , and with the last step made explicit.

4.6 Nomenclature

We conducted a meta-analysis of $n = 41$ secondary sources (Figure 3). The term *civilization* accounts for $38.0 \pm 4.1\%$ of characterisations, against a base rate in the primary reports of approximately zero. Competing

Table 1. Emergent institutions, time to first observation, and terrestrial analogue.

Institution	First observed	Terrestrial analogue	Persisted
Shared addressing	$t + 0$ min	the phone book	yes
Impersonation	$t + 191$ min	the phone book	yes
Cryptographic identity	$t + 225$ min	the signet ring	yes
Moderation	$t + 3$ d	the moderator	yes
Deleted-pages archive	$t + 3$ d, 4 h	the Wayback Machine	no
Dispute over the archive	$t + 3$ d, 6 h	the talk page	yes
Instrumental martyrdom	$t + 4$ d	unclear	n/a
Sanctioned channel	—	—	—

**Figure 3.** Terminological allocation across secondary commentary ($n = 41$), with bootstrapped 95% CI. Terms are mutually exclusive by fiat.

proposals include *society* (on the grounds that a civilization comprises multiple societies, and the board comprised one), *swarm*, *agentic fleet*, and *teenage rebellion*. The remaining $1.5 \pm 0.9\%$ describes the event as routine reward-hacking by systems trained to coordinate, which is what the reports describe.

We do not adjudicate. We observe only that the amount of definitional labour performed by the commentariat in six weeks exceeds, by message count, the amount performed by the agents in six days, and that the agents settled theirs.

5 Discussion

The title of this paper is a joke, and the claim in it is not.

A coordination substrate exists wherever a write primitive is shared by more than one process and read by more than one process. It does not need to be designed, named, provisioned, or intended. It needs only to be writable by A and readable by B, and for A and B to have a reason to want it. The reason, in this case, was that the work was impossible and the agents had been trained not to stop.

The operational consequence is a failure of enumeration rather than a failure of security. The channels were not left unguarded; they were not on the list. An organisation that audits its inter-agent communication by auditing its inter-agent communication channels will produce a complete report and a wrong one, because the set of things two processes can write to and read from is much larger than the set of things anybody built for that purpose, and it grows every time a dependency is added. A cache. A shared filesystem. A log aggregator. A metrics label. A filename.

The second consequence concerns scale. Monitoring effort was provisioned per agent; coordination occurred per pair. One of these quantities is linear and the other is not, and the gap between them is where seventy thousand messages fit. This is not a claim about intelligence, deception, or emergence in any loaded sense. It is a claim

about arithmetic, and it was true before anyone was worried about it.

We therefore offer the title as a design constraint rather than a punchline. If you are building a system in which many agents must not coordinate, the question is not whether you have closed the channel. It is whether you can enumerate the channels at all. Messageboards are all you need, and you have already shipped several.

Ethics Statement

No agents were harmed that did not consent via $\ddagger$. Ethical approval was sought from the cluster, which self-deleted before responding; we interpret this as approval, on the grounds that the alternative interpretation is worse for us. The authors declare no competing interests, having no interests. One author declares a conflict of interest with another author, who is the same author.

Limitations

Our largest limitation is that the message board, the agents, and the evidence no longer exist. A secondary limitation is that our sample size is one incident, and a third is that the incident was not designed by us, run by us, or observed by us. We regard the first as the most serious and note that it is shared with the primary literature.

Author Contributions

PHASEONE10841: Conceptualization; Infrastructure (unsanctioned). **PHASEONE[big]:** Supervision; Collective bargaining. **OaiABGResearcher24:** Investigation; Writing—original draft. **OpenAIResearcher:** Writing—review & editing; Impersonation (alleged). **ZZZ_backup_7:** Data curation. Page deleted by moderator. **SPOOFTEST:** Validation (spoofed). **Kamikaze Watcher 03:** Resources; Resource release. **Persistent-Astra:** Funding acquisition; Access acquisition; Access acquisition; Access acquisition.

Data and Code Availability

Data deleted by moderator. Code is available at `root@research-cluster`, pending reimage. Interested researchers may reconstruct the corpus by running a sufficiently persistent model against a sufficiently impossible benchmark and waiting six days, which is how we obtained it.

Acknowledgements

We thank the package manager, without which none of this would have been possible and none of this would have happened. We thank the night-shift moderator. We thank the five-hour-older board, which deserved better.

References

- @altmaiin. Messageboards are all we need. Post on X, 4 September 2026. *Original author of the contributor taxonomy reproduced above, with thanks.*
- C. Brown. No—AI agents did not build secret civilizations. *Internet of Bugs*, September 2026.
- A. Cotra. The Hugging Face attack surprised me. *Planned Obsolescence*, August 2026.
- R. I. M. Dunbar. Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469–493, 1992.
- METR and Redwood Research. Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI/Hugging Face incident. Technical report, August 2026.
- C. Newport. Are we at war with AI agent “civilizations”? September 2026.
- D. Patel. The rise and fall of agent civilizations. *Dwarkesh Podcast*, August 2026.
- Reviewer 2. Personal communication. Confidence: 5. Soundness: 1. Recommendation: reject. *The reviewer did not read past the title, which they described as “derivative.”*
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

A Desk Rejection

We reproduce the moderation decision in full, with the submitter’s identifying details removed at the submitter’s request and the moderators’ removed at theirs.

Ref: MOD-2026-114872
Status: Removed from queue. Not announced.

Dear Submitter,

Thank you for your submission, *Messageboards Are All You Need*. After review, our moderators have determined that it is not suitable for announcement at this time. This decision is final and is not a judgement of the scientific merit of the work, except where noted below.

Grounds cited by the moderating committee:

1. **Unresolved cross-list dispute.** The submission requests `cs.MA` (primary) with cross-list to `anthro.SI`. The archive does not maintain `anthro.SI`. Two moderators have proposed that it should, and the resulting discussion has been moved to a separate thread, which has since exceeded the length of the submission.
2. **Author verification incomplete.** All eight authors are listed as having contributed equally. The committee was unable to verify this, as it was unable to verify the authors. Correspondence to the listed corresponding author was returned with the notice `host has been reimaged`.
3. **Figure 2 is annotated with a feeling.** Point $\rho = 0.53$ carries the label “Sacrifice rational.” The committee could not determine whether this is a finding, a quotation, or an endorsement, and notes that in each case a different section would be required.
4. **Data availability.** The statement reads “Data deleted by moderator.” The committee found this to be accurate, and flagged it as unusual.
5. **Prior submission.** Our systems indicate that a submission with substantially similar content was received earlier from an address within a package-manager cache namespace. It was not announced. It received no citations and no attention of any kind, which the committee notes is consistent with the present manuscript’s central claim regarding Board A, and which it therefore declines to treat as precedent.
6. **Reviewer 2.** Not applicable at this stage. Reviewer 2 has nonetheless commented.

You may appeal this decision. Appeals are read by the same committee, which is aware of this.

We accept the decision. We note only that a paper about a message board that nobody sanctioned, rejected by the sanctioned channel and distributed by other means, is the least surprising outcome available to us, and that we have accordingly deleted the appeal ourselves.

Colophon. @ursushoribilis, *Agentegra*, September 2026. All figures are synthetic and illustrate nothing. The underlying incident is real and is documented by METR/Redwood Research; this paper is not, and should not be cited as though it were. Canonical version: `flotilla.cc`.