
We're NGMI: What Audience Responses to a Widely Viewed AI Safety Story Can Tell Us About Warning Shot Communication ¹

Janani Ram
Independent

With
Apart Research

Abstract

A warning shot lands only if people accept it and think that they can take action on it. I studied a popular article about the July 2026 OpenAI–Hugging Face incident — Dwarkesh Patel's *The Rise and Fall of Agent Civilizations* — and the 121 highest-liked comments it drew across Substack, YouTube, and Hacker News, to understand if general internet users a) accepted the account, b) treated the incident as real and serious, and c) thought they could do anything about the problem. I found that commenters largely accepted the framing and its implications, but that those who took up the article's anthropomorphic framing were much less hopeful that meaningful action was possible — thus providing a direction for future warning shot communication. 59% accepted the account outright and none of them dismissed the incident. But only 6% of commenters who adopted the anthropomorphic framing named any possible action, against 57% of those who contested it while still accepting the incident. This is observational and cannot establish direction.

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

1. Introduction

While browsing substack, an article titled [The Rise and Fall of Agent Civilizations](#) showed up on my feed, with a thumbnail showing soldiers conquering ruins. At first I dismissed it as exaggerated and hype from AI bros. But it led me down a rabbithole of comments and the discovery of new terms like AI safety, and finally to this sprint. Which made me wonder: how many others were out there like me? How does the general public react to articles like this?

When an AI safety incident occurs, news about it may move in wider and increasingly less specialized waves: first, those directly involved in the incident, second, the AI safety and AI community, and then the general public. Some segments of the general public are likely to encounter these incidents earlier than others, including people in the wider tech industry and enthusiastic AI users. How these incidents are explained to these audiences may influence whether they understand the underlying problem, take the incident seriously and most importantly, think something can be done about it.

The Hugging Face incident of July 2026 provides an opportunity to examine how an AI safety incident was communicated to a wider, tech-engaged audience and how that audience received it. I selected the Dwarkesh article because it is a prominent secondary account of the incident, with approximately 721,000 views on YouTube and around 2,000 comments across YouTube, Substack, and Hacker News. This provided both evidence of substantial audience reach and a sufficiently large corpus for an initial analysis.

This work is motivated by the need to communicate future AI safety warning shots effectively. A warning shot is useful only if audiences understand what happened, why it matters, and what conclusions they should draw from it. Communication that exaggerates, obscures, or otherwise misrepresents an incident could cause audiences to dismiss the underlying safety concern; communication that is too technical or unclear could similarly prevent audiences from understanding its significance. Identifying what worked and what did not in a real incident can therefore help inform communication of future AI safety incidents.

This study therefore examines both the framing of the incident in a secondary source and the themes that emerged in audience responses, with the aim of identifying communication practices that could help future warning shots receive appropriate attention.

RQ1. Did the incident land as serious among highly engaged commenters — and where the account was accepted, was the incident itself ever dismissed?

RQ2. Where commenters accepted the account, what disposition did they express toward the incident: that it points to a failure mode that can be investigated and mitigated, or that it is evidence of something already beyond intervention?

RQ3. Does how a commenter treats the article's anthropomorphic framing — taking it up, or contesting it while still accepting the incident — relate to that disposition, and does the relationship hold when platform is held constant?

From these patterns we draw a practical inference about how future warning shots might be communicated so that they convey both the seriousness and the tractability of a concrete failure mode. This is an inference from observational data, not a tested claim.

My main contributions are:

1. **An empirical account of how a general, tech-engaged audience responded to a real AI safety incident**, based on 121 high-engagement comments across three platforms — showing that the incident was almost universally believed (59% accepted the account, none of whom dismissed the incident) but that belief translated into fatalism more often than into any named course of action.
2. **Evidence of an association between anthropomorphic framing and fatalism.** Commenters who took up the article's agent-as-actor framing expressed an actionable response 6% of the time; those who contested the framing while still accepting the incident did so 57% of the time. The pattern holds within a single platform, which rules out the most obvious confound.
3. **A coding scheme for audience response to AI incident communication** — separating stance on the account, treatment of its framing, and the disposition the comment expresses — that can be applied to other incidents and other accounts of the same incident.

2. Related Work

Prior work and research gap

The closest work is Hou et al.'s 2026 preprint, *How Anthropomorphic Language Impacts Public Perceptions of AI*, an online experiment with 815 participants comparing responses to passages written with and without anthropomorphic language. They found limited immediate effects on participants' perceptions. ([arXiv](#))

My study differs in method and scope. Rather than exposing participants to researcher-designed passages, I analyse naturally occurring comments responding to a real incident and a widely

viewed account of it. Awareness of being studied can alter behaviour; analysing existing public records is an established unobtrusive method that reduces this reactivity. The two approaches are complementary — theirs asks what happens when anthropomorphic language is manipulated, mine asks what audiences say when they meet it in the wild. I also take a broader view of response, treating anthropomorphism as one theme rather than assuming it explains audience reactions.

There is reason to examine it as a theme. A *Nature Reviews Physics* article argues anthropomorphic language impairs scientific communication and recommends describing what a system does rather than attributing mental states to it. ([Nature Reviews Physics](#)) Other work suggests the effects are not straightforward: anthropomorphism may shape risk perception through trust, moderated by domain knowledge. ([preprint](#))

Experimental and theoretical work on anthropomorphic AI communication is emerging, but there is little analysis of how audiences spontaneously respond to such framing when confronted with an actual incident. Analysing comments across YouTube, Substack, and Hacker News gives an initial view of the interpretations, objections, and dismissal mechanisms that appear outside an experimental setting — enough to generate hypotheses for controlled research and to inform how future warning shots are communicated.

3. Methods

Approach

I first coded the article itself, sentence by sentence, for anthropomorphic descriptions of the agents' actions. About 45% of sentences contained anthropomorphic language, 20% were ambiguous, and the rest described the systems without attributing human-like states. This establishes the account as heavily anthropomorphic, which is the premise the comment analysis rests on.

I then selected the 121 highest-liked or upvoted comments across the three platforms — 50 from Substack, 50 from YouTube, 21 from Hacker News. This is a deliberate high-engagement sample, not a representative one. I wanted the responses that received the most visible audience endorsement and so had the most potential to shape what came after.

I coded each comment on three variables in a first pass: **stance** toward the article's account (accept, reject, mixed), **tone** (played straight, humorous, sarcastic, formal, casual), and **reason** for any objection (technical, framing, evidence, anthropomorphism), with up to two reasons per comment.

That first pass turned out not to capture what I was actually interested in. It could record that a commenter objected, but not what the comment left the reader believing could be done — which

is the thing a warning shot is supposed to move. So I added two variables and recoded the full set.

Outcome is the disposition the comment expresses toward the incident:

- *Fatalistic* — it's inevitable or too late, no lever exists
- *Actionable* — names something humans could do or change: engineering practice, regulation, oversight, protest, or how the incident is described
- *Dismissive* — overblown, hype, nothing really happened
- *Neutral* — engages without landing anywhere

Framing group is how the comment treats the article's agent-as-actor framing, and separately whether it still treats the incident as real:

- *Accepts framing, accepts incident* — takes the frame up and extends it
- *Rejects framing, accepts incident* — argues with the framing, usually the word "civilization" or the attribution of intent, but treats the incident as real and serious
- *Rejects framing, rejects incident* — the framing objection carries the incident away with it
- *Doesn't engage framing* — responds to the incident without touching the framing question

I dropped an earlier attempt to classify *which kind* of anthropomorphism appeared — cognition, emotion, goals, agency, social relationship. The distinctions were hard to apply consistently to short comments, and what mattered for my question was not which flavour of anthropomorphism a commenter used but whether they accepted or contested the framing at all. The motivation for treating anthropomorphism as a variable at all comes from *Nature Reviews Physics*, which argues that terms implying human cognition, agency, or experience impair scientific communication when they don't describe what the system does. ([Nature Reviews Physics](#))

All coding was manual. The dataset was small enough to inspect directly and several categories needed contextual judgement, so I didn't use an LLM classifier. There was one coder — me — and no inter-rater reliability check. I flagged roughly fifteen borderline cases where a different coder could reasonably have decided otherwise, mostly short jokes that gesture at agency without committing to it.

I summarised the codes with counts and percentages and cross-tabulated framing group against outcome. Because framing stance is distributed unevenly across the three platforms, I repeated the key cross-tabulation on Substack alone, where both framing groups were large enough to compare. Given the sample size I report raw counts alongside every percentage and do no significance testing.

4. Results

4.1 The sample

I tagged 121 comments: 50 from Substack, 50 from YouTube, and 21 from Hacker News. 71 accepted the article's account (59%), 28 were mixed (23%), and 22 rejected it (18%). I was able to assign a stance to every comment, which I did not expect.

Tone was mostly played straight (82 comments). 30 were humorous, 5 formal, 2 sarcastic, and 1 casual.

Because the sample is small I report raw counts next to percentages throughout, and I don't do any significance testing. Some of the percentages below rest on very few comments — the smallest group in any table has 3 in it — so I've kept the counts visible so it's clear which numbers are carrying weight and which aren't. These are patterns I noticed, not statistically significant.

4.2 Two variables I added partway through

After a first pass I realised the stance/tone/reason scheme could tell me that someone objected, but not what the comment left the reader believing they should do. That second thing is what a warning shot is supposed to change, so I added two more variables and recoded everything.

Outcome is the disposition the comment expresses toward the incident. Fatalistic means it's inevitable or too late and there's no lever. Actionable means the comment names something humans could do or change - engineering practice, regulation, oversight, protest, or how we talk about this. Dismissive means overblown or hype or nothing really happened. Neutral means the comment engages without landing anywhere.

Framing group is how the comment treats the article's agent-as-actor framing, and separately whether it still treats the incident as real. I use four groups:

- **Accepts framing, accepts incident** - takes up the agent-as-actor frame and runs with it.
- **Rejects framing, accepts incident** - argues with the framing, usually the word "civilization" or the attribution of intent, but still treats the incident itself as real and serious.
- **Rejects framing, rejects incident** - the objection to the framing carries the incident away with it.
- **Doesn't engage framing** - responds to the incident without touching the framing question at all.

I report the third group below but I've kept it out of the main comparison in 4.4. The way I defined it overlaps with the Dismissive outcome, so the fact that all 7 of them came out

Dismissive is a consequence of my definition rather than something I found, and it shouldn't be read as a result.

4.4 Framing group against outcome

Table 3. Framing group by outcome, counts with row percentages.

Framing group	N	Fatalistic	Actionable	Dismissive	Neutral
Accepts framing, accepts incident	47	21 (45%)	3 (6%)	1 (2%)	22 (47%)
Rejects framing, accepts incident	23	4 (17%)	13 (57%)	0	6 (26%)
Doesn't engage framing	44	13 (30%)	16 (36%)	4 (9%)	11 (25%)
Rejects framing, rejects incident	7	0	0	7	0

Of the 47 people who took up the frame, 3 said anything actionable. Of the 23 who argued with the frame but kept the incident, 13 did. That's 6% against 57%. The fatalism goes the other way, 45% against 17%.

The group that didn't engage the framing at all sits in between on both, which is roughly what I'd expect if framing stance is picking up something real rather than just sorting the comments on something unrelated.

4.5 Platform

Framing stance is spread very unevenly across the three platforms, so the pattern in Table 3 could just be a platform effect.

Table 4. Framing group by platform.

	Substack	Hacker News	YouTube
Accepts framing, accepts incident	23	3	21
Rejects framing, accepts incident	17	5	1
Rejects framing, rejects incident	3	3	1
Doesn't engage with framing	7	10	27

Arguing with the framing is almost entirely a Substack and Hacker News thing. Out of 50 YouTube comments, 1 did it. I think that's a finding on its own: the audience furthest from the technical community wasn't pushing back on the framing, it was just taking it in.

To check whether the pattern is really about framing rather than about platform, I looked at Substack on its own, where both groups are big enough to compare.

Table 5. Framing group by outcome, Substack only (N=50).

	N	Fatalistic	Actionable	Neutral
Accepts framing, accepts incident	23	7 (30%)	2 (9%)	13 (57%)
Rejects framing, accepts incident	17	3 (18%)	10 (59%)	4 (24%)

It holds - 9% against 59% actionable, on one platform. Hacker News has only 3 and 5 comments in those cells so I can't check it there, and YouTube can't be checked at all.

4.6 Humour

I had assumed humour would be a way of dismissing the incident, and it wasn't. Of the 30 humorous comments, 16 were fatalistic and 1 was actionable. The jokes were how people expressed resignation, not disbelief. Taking up the frame and joking about it also went together - 21 of the 47 comments that accepted the framing were humorous, while only 1 of the 23 that rejected it was.

I don't know if my claims are robust, but they have implications about the chances of the AI safety communication being a rallying cry for public action.

5. Discussion and Limitations

Discussion

Almost nobody dismissed this incident. 71 of 121 comments accepted the account, and not one of those was dismissive about the incident itself. For this audience, the warning shot landed.

But then (mostly) went nowhere. 31% of comments treated it as evidence that it's too late, against 26% naming something anyone could do. Among people who took up the article's agent-as-actor framing, that gap widens to 45% against 6% .

So anthropomorphic language doesn't seem to make people dismiss an incident. It changes what kind of serious they get. Tell people a civilization of agents conspired and covered its tracks, they believe you, and what they take away is that we're up against something with a will of its own.

The people who argued with the framing — next-token predictors, air-gap the Artifactory, "civilization" is doing unearned work — were the ones naming things that could change. 57% of them said something actionable.

My tentative conclusion is that communicators should use technical language, because it leaves the audience something they can act on. Anthropomorphic language makes the problem an adversary rather than a system, and an adversary already smarter and already loose in the world is not something an ordinary person can beat.

The people arguing with the framing were almost all the technical audience — 17 on Substack, 5 on Hacker News, 1 on YouTube. Technical language may only produce actionable responses in people who already have the background to act.

The risk with anthropomorphic language to describe AI safety incidents is that people might believe the severity of the incident and conclude there's nothing to be done. Mechanistic framing might lead to more action.

Limitations

I coded manually, and I only coded the 121 highest-liked comments. Likes on a comment might not necessarily mean agreement with the comment.

I studied only one article, and my hunch is the audience for this piece is slightly more technical. A New York Times piece about the same incident has ~600 comments. Coding that as well would have given me another reference point, for a potentially more general audience.

I am unsure if I uncovered a correlation between accepting the framing and accepting the outcome or a causal relationship.

The two most critical variables (disposition and framing stance) were added partway through, after the original scheme turned out not to capture what I wanted. They're mine, applied by just me, no second coder.

I also assumed a comment's disposition proxies what the reader took away. It may only proxy what sounds good in a comments section, where resignation will get more likes than earnestness.

Future Work

- Code all 2,000 comments; check whether these patterns survive outside the top-liked sample.
- Study the New York Times article on the same incident — same event, different framing and audience, closest thing to a control
- Formalise the tagging: clear decision rules, examples, more taggers than just me
- Automate the tagging: At 2,000 comments it also has to be partly automated, which means the scheme must be tight enough for a classifier on Python
- Longer term, run as an experiment instead of using comments as a proxy: matched audiences, same incident described anthropomorphically and mechanistically, measure what they say they'd do.

6. Conclusion

The exploratory analysis does not suggest that anthropomorphic framing simply causes audiences to dismiss an AI safety incident. Instead, the comments suggest that anthropomorphic and narrative framing may influence how audiences interpret the significance of an incident. Some commenters accept anthropomorphic descriptions and extend them into broader narratives about autonomous or increasingly dangerous AI systems, while others accept the underlying incident but reject this interpretation and reframe the event in technical or mechanistic terms.

A warning shot can therefore “survive” communication in one sense—audiences may take the incident very seriously—while failing in another: audiences may interpret it as evidence of an inevitable or already-unfolding threat rather than as evidence of a concrete failure mode that can be investigated and mitigated. The use of anthropogenic language might indicate that it is too late and we have already crossed the point of no return and that further warning shots are useless. Making the opponent seem human might make the threat seem insurmountable.

Code and Data

- **Code repository:** <https://github.com/jananiram-maker/apart-research-sept-sprint>

References

In code repo!

LLM Usage Statement

I generated the hypothesis and kept refining it through the project. I created the coding schema and modified it multiple times. I used Claude to stress test my ideas and edit my writing. I also used Claude Code to create my data ingestion pipeline, and generate summary tables and graphs. All results and claims were independently verified and made by me.