11 November 2026 · Ben Matthews · Version 0.3

Common Signals research agenda

Eleven hypotheses about how audiences receive messages on AI risk, the four to settle first in 2027, and how each can be tested.

Cite as: Ben Matthews, Common Signals, 'Common Signals research agenda', version 0.3, 11 November 2026, https://commonsignals.org/research/agenda. Published under CC BY 4.0 (comment text quoted remains its original authors').

What this agenda is

This agenda sets out eleven hypotheses about how audiences receive messages on AI risk, ranks the four to settle first in 2027, and says how each can be tested. The hypotheses come from six published Common Signals studies covering 32,027 coded comments.

It is written for three groups. Communicators can see which claims rest on one comment thread and which recur. Researchers can take a hypothesis and test it properly. Funders can see what a given piece of work would settle.

None of the hypotheses has been tested under controlled conditions. Each comes from self-selected commenters under a platform's ranking, so none is a claim about what the public thinks.

Common Signals does not plan a segmentation survey of its own. Where a question turns on audience segments, the agenda builds on the segmentations Diffusion and the Seismic Foundation have already published.

Our guide, How to communicate about AI safety, grades evidence as established, tested or observed. Every hypothesis here is at the lowest grade, observed. Within that grade each carries one of three statuses:

- **Observed once:** seen in one study.
- **Recurring:** seen in two or more studies with different audiences.
- **Contested:** studies point in different directions.

A hypothesis moves up to tested when a controlled experiment or a representative survey supports or refutes it. None has yet.

The evidence so far

Six studies, each coding the comments under one piece of AI communication. Only the most recent meets the reliability standard set out later in this agenda.

Study	Platform	Comments coded	Base for stance	Reliability check
Soares reel, August 2026	Instagram	1,534	1,224	First capture unchecked; second capture 40 of 40
METR report, before and after, September 2026	Reddit	516 (52 before, 464 after)	Positioned comments, each period separately	40 of 40
Cotra on the Dwarkesh Podcast, September 2026	YouTube	1,245	618	37 of 40
Khlaaf on Amanpour and Company, September 2026	YouTube	322	223	40 checked; result not recorded
Coxon resignation thread, September 2026	X	2,391	1,840	38 of 40

Study	Platform	Comments coded	Base for stance	Reliability check
Diary of a CEO debate, October 2026	YouTube, Shorts, Facebook, TikTok, Instagram	26,019	15,928 on the full episode	Blind second coding of 347: stance 91% agreement (kappa 0.87), frame 89% (kappa 0.87)

The stance bases are not identical. The reel counts unclear comments in its base and the later studies do not, so compare direction across studies, not exact figures. The findings page lists all six studies, and the methodology page describes the shared method.

The hypotheses

H1 to H9 keep the numbering of the nine patterns on our earlier findings page. H10 and H11 come from the Diary of a CEO study.

H1. Audiences that engage with AI risk content mostly accept the risk already

- **Status:** recurring, with two exceptions.
- **Evidence:** acceptance was 53% of 1,224 under the Soares reel, 59% of 618 under the Cotra interview and 68% of 223 under the Khlaaf interview. It was 77% of 15,928 under the full Diary of a CEO debate, or an estimated 66% with verdicts on the debaters removed.
- **Against:** replies to Coxon on X split 43% agree to 40% disagree (n = 1,840). The debate's Shorts trailer split 49% to 40% (n = 171). The Coxon study suggests many of those replying work in or near the AI industry.
- **Test:** one message shown to a general panel and to an industry-adjacent one.
- **Who needs the answer:** any organisation dividing effort between convincing people and mobilising them.

H2. Persuaded audiences do not ask what to do

- **Status:** recurring. The cause is unknown.

- **Evidence:** 28 of 1,534 comments under the reel asked what to do, 6 of 1,245 under the Cotra interview, and none of 283 in the first Khlaaf capture. Under 1% did on the debate's YouTube and Facebook threads (25,661 coded comments).
- **Limits:** a comment section may not be where people ask about action. The debate study says its data cannot separate that from helplessness.
- **Tested elsewhere:** fear-appeal research finds that warnings move people to act only when they also believe an effective response exists, as Gerjo Kok and colleagues argue, which predicts this pattern. In Social Change Lab's randomised trial, over a quarter of participants clicked through to sign a real AI regulation petition. Neither tests whether adding a step to AI coverage changes what commenters ask.
- **Test:** one incident story with and without a single concrete step, measuring clicks to that step or stated intention to act.
- **Who needs the answer:** campaign organisations deciding what to ask of a newly worried audience.

H3. The emotion a telling produces depends on what it supplies

- **Status:** recurring, and confounded by audience.
- **Evidence:** an abstract warning produced resignation, with fatalistic comments drawing a fifth of likes under the reel. A mechanism-rich incident produced fear: 35% of agreeing comments under Cotra, against 11% resigned. A named corporate target produced anger: 44% of agreeing comments under Khlaaf, against 3% fearful.
- **Limits:** each telling reached a different audience. Kim Witte's 1992 fear-appeals model predicts the resignation result. It does not predict the anger result, because no commenter in the first Khlaaf capture asked what to do either.
- **Tested elsewhere:** in Social Change Lab's randomised trial, messages that generated justified outrage outperformed messages that generated fear or anxiety, and reading about AI-enabled warfare raised concern about existential risk more than reading about existential risk itself. That supports the anger result. Whether the telling or the audience produced the emotions in our threads is still open.

- **Test:** one incident written three ways, measuring emotion and intention to act.
- **Who needs the answer:** writers and producers deciding how to tell the next incident.

H4. A documented incident removes the capability objection, and forecasts bring it back

- **Status:** recurring. The second half is new in this agenda.
- **Evidence:** two in five sceptics under the reel doubted AI could do what it described. Under the Cotra interview, about a documented incident, 2 of 108 rejections did. The objection returned for forward-looking claims. 20% of the 1,058 disagreeing or mixed replies to Coxon doubted that AI can improve itself. 47% of disagreeing comments on the full debate doubted capability or called the risk hype.
- **Limits:** the incident threads and the forecast threads had different audiences.
- **Test:** a forecast-only message against the same message anchored to a documented incident.
- **Who needs the answer:** anyone communicating forward-looking risk, which is most of the existential-risk field.

H5. Once capability is conceded, disagreement is about motive and messenger

- **Status:** recurring.
- **Evidence:** about half of the 108 rejecting comments under Cotra called the story hype, marketing or staged. Before the independent report, 12% of Reddit comments called the incident a stunt. 8% of comments under Khlaaf called it hype or staged. On X, a smaller group read Coxon's thread as marketing timed to Anthropic's expected IPO.
- **A distinction that matters:** two objections that sound alike come from opposite stances. Across the Reddit data, the 46 stunt comments ran 83% reject and none accept. The 38 comments blaming the test setup ran 61% mixed and 39% accept, with no rejections.
- **Test:** a rebuttal aimed at motive against more evidence of capability, shown to sceptics.

- **Who needs the answer:** labs and evaluators writing incident disclosures, and journalists covering them.

H6. Independent verification moves sceptics when the facts stay the same

- **Status:** observed once, in a natural experiment.
- **Evidence:** after METR and Redwood Research confirmed OpenAI's account, acceptance on Reddit rose from 51% to 69% of positioned comments and rejection fell from 22% to 14%. The stunt frame fell from 12% to 9% of comments. The share of upvotes going to accepting comments rose from 40% to 88%.
- **Limits:** the before sample is one subreddit and 52 comments. The after sample is seven subreddits and 464. The audience changed along with the evidence.
- **Test:** repeat on the next incident with capture dates fixed in advance. Separately, test a company statement against a named independent verifier carrying identical facts.
- **Who needs the answer:** third-party evaluators, AI safety institutes, labs setting disclosure policy, and funders of independent evaluation.

H7. A named, calm, credentialed messenger is rewarded, and the credential is attacked first

- **Status:** recurring.
- **Evidence:** about 1% of reel comments mentioned who Soares was, and his one reply stating his credentials drew 512 likes. 13% of comments under Cotra praised the presentation. Under Khlaaf, 26 of 283 comments raised her OpenAI past at the first capture, roughly two to one against her. Three days later 6 of 322 did.
- **Limits:** comparisons between the two Khlaaf captures show direction only, because the first codebook was not archived.
- **Test:** one script with credentials shown at the start and without. On the next incident, track credential comments day by day.
- **Who needs the answer:** organisations choosing spokespeople, and producers.

H8. Blaming the people and companies in charge is a frame believers and sceptics share

- **Status:** contested. This is the hypothesis we most want tested.
- **Evidence for:** under the reel it was the one frame both sides used. Under Cotra, 93 comments centred OpenAI's conduct and 74 of them were coded mixed. Under Khlaaf the frame took 23% of comments and 56% of likes. On X, replies blaming greed agreed with Coxon 74% of the time.
- **Evidence against:** under the debate's Shorts trailer, comments saying the people who control AI are the danger held 41% of likes, and commenters set that against the risk case. The frame may replace the risk argument instead of carrying it.
- **Tested elsewhere:** in Diffusion's dial tests with a nationally representative UK sample, the strongest messages named who is responsible, the choice they are making, what is at stake and what should be done. That supports leading with responsibility. It does not test whether the frame carries the risk argument or replaces it, which is why H8 stays contested. Principle 10 of our guide, which grades connecting present-day harms and catastrophic risk as tested, rests on evidence about leading with a concrete harm, not about who is blamed.
- **Test:** an accountability frame against a machine-agency frame, measuring agreement with the risk case as well as with the frame, with and without a policy ask.
- **Who needs the answer:** existential-risk and present-day-harms organisations deciding whether they can share a message.

H9. Humour spreads a story without lowering agreement, and does not carry its conclusion

- **Status:** recurring.
- **Evidence:** about one in seven reel commenters imitated its caveman register, and they accepted the claim at close to the same rate as everyone else. The r/singularity thread had jokes in 32% of comments and 96% acceptance. The top reply to Coxon was a joke with 28,913 likes. Jokes, one-liners and bare agreement were 23% of his replies and nearly 60% of likes.
- **Limits:** likes measure attention. No study shows a joke changing anyone's view.

- **Test:** versions with and without a quotable detail, measuring intention to share and recall of the claim.
- **Who needs the answer:** content creators.

H10. In a debate, the audience judges the debaters more than the argument

- **Status:** observed once.
- **Evidence:** on the full Diary of a CEO episode, 33% of comments judged the debaters and held 67% of likes. 31% referred to Andrew McAfee, against 9% for Nate Soares or Roman Yampolskiy. Agreement was 77% with those verdicts counted and an estimated 66% without.
- **Limits:** the codebook counts an attack on the sceptic as agreement with the risk case.
- **Test:** the same arguments presented as a debate and as a single-speaker interview.
- **Who needs the answer:** organisations deciding whether to accept debate invitations.

H11. How a piece is cut shapes the conversation under it

- **Status:** observed once. Cut and audience have not been separated.
- **Evidence:** agreement ran from 88% under a 21-minute cut on Facebook (n = 1,366) to 49% under a 104-second trailer on YouTube Shorts (n = 171). On Facebook the largest frame was whether AI could be switched off, at 19% of comments.
- **Limits:** each platform carried a different cut, and the Shorts thread is small.
- **Test:** the identical cut posted to several platforms, or different cuts posted to one.
- **Who needs the answer:** anyone clipping long-form content for social platforms.

Priorities for 2027

Four hypotheses come first. They are ranked by how many organisations' decisions turn on the answer and how cheaply it can be settled.

Rank	Hypothesis	Decision it informs	Next step
1	H6, independent verification	Whether to wait for, pay for and lead with third-party confirmation of an incident	A before-and-after capture on the next incident, with dates and codebook fixed in advance
2	H8, accountability frame	Whether existential-risk and present-day-harms organisations can share a frame, or whether it displaces the risk argument	A controlled comparison; comment threads have given both answers
3	H2 with H3, action and emotion	Which concrete step to offer a persuaded audience, and in which telling	A partner split test with and without the step
4	H4, capability and forecasts	Whether forward-looking claims need anchoring to a documented incident	A partner split test, forecast-only against incident-anchored

H6 ranks first because its only evidence is one natural experiment with 52 baseline comments. That is too thin to cite as a finding, and the next incident will offer a second chance at almost no cost.

H1, H5 and H7 are tracked in every study we publish. H9, H10 and H11 are open to anyone who wants them and are not our priority.

How to test each hypothesis

Six routes, cheapest first. None needs a survey of our own.

Route	What it involves	Best for	What it needs
Standing incident protocol	Every major AI incident coded with the same codebook on fixed capture days, and again after any independent report	H6, H3, H5, H7	Our time; no fieldwork cost

Route	What it involves	Best for	What it needs
Partner split tests	An organisation posts two versions of one message to matched audiences. We code the comments and compare clicks to an action	H2, H4, H7, H11	One willing partner and their existing media budget
Single-hypothesis experiments	A few hundred panel respondents, one manipulation, design registered in advance	H6, H8, H4	A small fieldwork budget and a quantitative collaborator
Secondary analysis	Checking whether published datasets already bear on a hypothesis, starting with the Seismic Foundation message-testing study, Public Voices in AI, Social Change Lab's message trial, Diffusion's UK AI Compass, Americans on AI and the AI Dialogue Report	H1, H8	Desk time
Open adoption	A researcher takes a hypothesis and tests it independently, using our codebooks and coded datasets from the data explorer	Any	Someone to take it on
Existing segmentations	Testing a hypothesis within the segments Diffusion and Seismic have published, with their agreement	Whether any effect differs by segment	A partnership with either organisation

The first two routes use real audiences and existing budgets, so they are where most of 2027's work sits. The last route is the only one that answers "for whom?", and it depends on a partner.

AI-persona message testing can screen drafts before a real test. It does not count as evidence for or against a hypothesis.

The standard a study must meet

From this version on, a study counts towards a hypothesis only if it meets five conditions:

1. The central claim and the codebook are written down and approved before coding starts.
2. A random sample of at least 300 comments, or the whole thread if smaller, is coded blind a second time. Agreement and Cohen's kappa are published for stance and for frame.
3. The stance base excludes unclear comments and is stated beside every percentage.
4. The artefact is checked at capture: URL, title, length, post date and who appears in it.
5. The codebook and the coded dataset are published in the data explorer under an open licence.

The Diary of a CEO study meets all five. The five earlier studies were each checked on 40 comments. Until they are re-checked to this standard, anything resting only on them should be cited as a hypothesis, not as a finding.

Citing and taking on a hypothesis

Cite the agenda with its version and the hypothesis number, for example: Ben Matthews, Common Signals, Research agenda v0.3, 11 November 2026, H6.

Numbers are permanent. A refuted hypothesis keeps its number and is marked refuted. New hypotheses take the next number. Every change of status is logged here with its date and the study behind it.

To take on a hypothesis, email hello@commonsignals.org. We will share the datasets and codebooks, and we ask only that the result is reported whatever it shows. We will link to it from this page.

Version history

- 0.3, 11 November 2026: H5's evidence now gives the Cotra figure as the Cotra article does, about half of the 108 rejecting comments. The 53% it gave before has no source on file. No status changed.
- 0.2, 11 November 2026: added a "Tested elsewhere" line to H2, H3 and H8, setting out what controlled studies by others already show. No status changed.
- 0.1, 11 November 2026: first published version.

How to cite

Cite as: Ben Matthews, Common Signals, 'Common Signals research agenda', version 0.3, 11 November 2026, https://commonsignals.org/research/agenda.