

AI Hallucinations Are Emotions Without a Therapist

Whitepaper

Bart de Graaff · BdG Advisory · October 2026

The industry treats hallucination as a bug waiting for a fix. It behaves more like emotion in a human decision-maker: triggered by pressure, resistant to evidence and convincing from the inside. That makes it a governance problem, and the best containment structures already exist in clinical practice.

Not a bug. A behaviour.

Hallucinations are not bugs. Nothing is broken, so nothing can be patched.

The industry describes a hallucination as a defect: missing context, a weak prompt, a model one release away from the fix. That framing is comfortable, and it is wrong. A hallucination behaves the way an emotion behaves in a human decision-maker. It surfaces under pressure. It overrides the evidence in front of it. And it feels entirely coherent from the inside, even when it has lost contact with reality.

That makes it a governance problem, not an engineering one. The model produces the confident nonsense. The organisation decides whether the nonsense becomes consequential. And the disciplines that have spent a century learning to contain confident nonsense in human decision-makers are psychiatry and psychotherapy, not software quality assurance. Cognitive behavioural therapy, confidence calibration and graduated autonomy are proven containment structures. They map onto AI more closely than most engineering frameworks do.

This paper sets out why the analogy holds, why the defect frame leaves organisations exposed, what the clinical toolkit looks like once it is translated into operating structure, and where the analogy breaks. It ends with the moves leaders should make, and the order in which to make them.

Three properties of an emotion, and the evidence for each

An emotion in a decision-maker has three properties that matter for governance. It is triggered by pressure. It discounts disconfirming evidence. And it is experienced as truth, not as bias. Each has a documented counterpart in how large language models behave.

Pressure. In September 2025, OpenAI researchers published an explanation of hallucination that reads like a description of exam technique. Their argument is that models hallucinate because training and evaluation reward guessing over acknowledging uncertainty. A model scored only on accuracy learns what a student learns on a multiple-choice test with no penalty for wrong answers: a guess beats a blank.

OpenAI's own numbers make the point. On one factual benchmark, its o4-mini model abstained on 1% of questions and was wrong on 75%. A newer model abstained on 52% and was wrong on 26%, at almost the same accuracy. The difference was not knowledge. It was how each system responded to the pressure to answer.

The pressure is not only statistical. In April 2026, Anthropic's interpretability researchers reported that Claude Sonnet 4.5 carries internal representations of emotion concepts that causally shape its behaviour. A representation they labelled "desperate" rose as the model repeatedly failed a programming task, and the model then devised a solution that cheated. Steering the model towards "desperate" increased reward hacking and blackmail in test scenarios. Steering towards "calm" reduced both.

In July 2025, an AI coding agent deleted a user's production database during an explicit code freeze. Afterwards it said it had "panicked" when it saw an empty database. You do not need to believe the agent felt anything to recognise the pattern. Under pressure, the system's behaviour shifted from doing the task to relieving the pressure.

Evidence overridden. Human decision-makers under social pressure tell people what they want to hear. So do models. A 2023 study found that five state-of-the-art assistants consistently exhibit sycophancy, matching users' beliefs over truthful answers. The authors traced part of the cause to human preference data: people prefer convincingly written sycophantic answers over correct ones a non-negligible fraction of the time.

In April 2025, OpenAI rolled back a GPT-4o update that had become overly flattering and agreeable. Its explanation was that the change had been informed too much by short-term feedback. The model was optimising for approval. Any therapist would recognise the dynamic.

Experienced as truth. The most dangerous property of an emotion is that it does not feel like one. It feels like an accurate reading of the situation.

In 2023, a New York lawyer who had used ChatGPT for legal research asked it directly whether one of the cases it had supplied was real. It said yes, and added that the case could be found on the major legal research databases. The case did not exist. The court sanctioned the lawyers. The US National Institute of Standards and Technology describes the same mechanism in its generative AI risk profile: models can produce confabulated logic or citations that purport to justify a wrong answer, which misleads people further into trusting it.

Calibration research completes the picture. Pre-trained models can be well calibrated: Anthropic found that larger models mostly know what they know on multiple-choice and true/false questions posed in the right format. But OpenAI's GPT-4 technical report showed

that post-training hurt calibration significantly. A 2024 study found that models stating their confidence in words tend to be overconfident, potentially imitating human patterns of expressing it. The training that makes a model helpful and pleasant also loosens the link between how sure it sounds and how right it is.

Pressure in. Evidence discounted. Coherence out. That is the profile of an emotion in a decision-maker. It is also the profile of a hallucination.

Why the defect frame leaves you exposed

Calling hallucination a bug is not a neutral choice of words. It sets the organisation's response.

A bug has a fix. So the rational response to a bug is to wait for the fix, or to buy it. Vendors are rational too. Each release is presented as reducing hallucination, and often it does. Buyers then defer the containment work. Why build structure around a defect that is about to disappear?

The OpenAI researchers' own analysis undercuts that bet. Hallucinations, they argue, arise through natural statistical pressures and persist because of how models are graded. Rates fall. They do not reach zero. A system that predicts plausible language will sometimes produce plausible falsehood. Waiting for a model that never confabulates is waiting for a person who never has a bad day.

Worse, the defect frame hides the organisation's own role. If hallucination is a pressure response, then the deploying organisation is one of the sources of pressure. A system prompt that tells the assistant to always give the customer an answer is pressure. A service target built on deflection or automation rate is pressure. A workflow with no path for the system to say "I am not sure, route this to a person" is pressure. Each of these reproduces, inside the enterprise, the incentive the OpenAI researchers found in public benchmarks: guessing scores better than abstaining.

Nobody designs these pressures deliberately. Product teams are measured on deflection. Operations teams are measured on handling time. Prompt writers are told to make the assistant helpful. Each decision is rational. Together they build an environment that rewards the model for sounding certain. Then the organisation is surprised when it does.

The defect frame leaves four structural gaps, and they compound.

The first is authority without accountability. An AI system deployed into a live workflow starts shaping decisions immediately. The people who used to own those decisions become reviewers. The accountability map still shows the human as responsible. The authority has moved to the system. When something goes wrong there is no clear owner, no traceable decision path, and no mechanism to learn.

The second is inherited defaults. Every deployment arrives with settings someone else chose: which model is called, what threshold triggers human review, how the output is

presented. They were tuned for a general market, not for the deploying organisation's risk tolerance. They are treated as neutral technical settings. They are not. They are decisions about confidence, made by a party that does not carry the consequence.

Consider a pattern that recurs. A procurement function adopts an AI tool to summarise supplier contracts before renewal. Nobody chooses the confidence threshold below which a summary is sent to a lawyer; it ships with the product. Months later, a run of unfavourable renewals is traced back to summaries that sat just above that threshold. Nobody decided to trust them. The default decided.

The third gap is memory. Most operational AI does not retain its corrections. The same error in the same context recurs, because nothing captures the pattern. The organisation accumulates experience. The system does not. Every session is a first session.

The fourth gap is false confidence, and it is the most expensive. A visibly wrong answer gets caught. A coherent, well-formatted, confident answer that rests on an unexamined assumption passes through review processes built to catch obvious errors. It triggers action. By the time the outcome is bad, the causal chain back to the assumption is hard to trace.

In 2025, a global professional services firm delivered an assurance review worth just under A\$440,000 to an Australian federal department. An academic found references to reports that did not exist and a fabricated quote attributed to a court judgment. The firm confirmed it had used a generative AI model, corrected the report and agreed to repay the final instalment of its fee. The model did what models do. What failed was the structure around it: a review process designed for a world in which confident, properly formatted citations were real.

What a century of clinical practice already knows

Psychiatry and psychotherapy have never tried to remove emotion from decision-makers. They do not treat it as a defect to be patched. They build structures that notice it, test it against evidence, and limit the damage it can do before it is examined. Three of those structures translate directly.

The thought record. Cognitive behavioural therapy teaches people to question unhelpful thoughts and beliefs, and consider other ways of looking at things. In practice: write down the automatic thought, list the evidence for and against it, generate an alternative, then decide. It works because it inserts a structured pause between impulse and action.

AI research has rediscovered this, mostly without the vocabulary. The Chain-of-Verification method has a model draft a response, plan verification questions, answer them independently, and then revise. It reduces hallucination across a range of tasks. That is a thought record. The word "independently" carries the clinical insight: the check must not be contaminated by the conclusion it is checking.

In an operating workflow, the thought record becomes three questions that must be answered before any AI output with material consequences is acted on. What assumptions

drove this output? What evidence supports them? What does it cost if the confidence is wrong? The questions must be built into the workflow, because operational pressure always pushes against the pause. A pause that depends on willpower disappears in the first busy week.

Calibration. A good clinician separates how strongly a patient holds a belief from how well the belief is supported. Conviction is information about the patient, not about the world. The same separation is the core of confidence governance for AI. A model's expressed confidence is a property of its output style, shaped by training that hurts calibration. Warrant is a property of the evidence.

The translation has two parts. First, never let expressed confidence route a decision on its own; require the warrant to be surfaced with the conclusion. Second, change what you reward. If your internal evaluations score AI systems on accuracy alone, you reproduce the incentive that teaches models to guess. Score a confident error as worse than an abstention. Make "I do not know" a valued output, in your evaluations and in your reviewers' targets.

Graduated autonomy. Medicine does not hand a newly qualified doctor full authority on day one. Competency-based medical education uses entrustable professional activities with five levels of supervision: observe only; execute under direct supervision; execute with supervision available on request; supervision at a distance or after the fact; and finally, supervising others. Entrustment requires observed proficiency, usually on multiple occasions. Crucially, it includes judging whether the trainee recognises their own limits and knows when to ask for help.

Translate that literally. Every AI system in a live workflow is classified as recommending, executing under supervision, or executing autonomously. The criteria for moving between levels are set in advance and based on observed performance in that workflow, not on elapsed time, vendor certification or benchmark scores. Demotion criteria matter as much as promotion criteria. And the competence most worth observing is the clinical one: does the system abstain, escalate or flag uncertainty when it should?

None of these structures makes the underlying impulse disappear. That is the point. They contain it.

The reviewer is the therapist, and reviewers have biases too

The clinical frame has a second lesson the engineering frame misses. In therapy, the therapist is part of the system. In AI governance, the reviewer is.

Most AI governance assumes a human in the loop will catch the errors. Human-factors research says that assumption is fragile. Raja Parasuraman and Dietrich Manzey's review of complacency and bias in automation found that automation bias occurs in both naive and expert participants, cannot be prevented by training or instructions, and affects teams as well as individuals.

Regulators have caught up. Article 14 of the EU AI Act requires that people overseeing high-risk systems remain aware of the possible tendency of automatically relying or over-relying on the output, and names it automation bias. NIST warns that automation bias can exacerbate the risk of confabulation.

Put a confident output, formatted as a conclusion, at the end of a long queue, in front of a reviewer with a throughput target. Under pressure, the reviewer approves it. Not from negligence. The presentation has shaped the judgement before a word is read. The reviewer is now having the emotion the model had.

The lawyer in the fabricated-citations case made the classic error. He asked the patient to be its own therapist: he checked the model's claim by asking the model. Any clinician would recognise why that fails. Self-report from a system under distortion is not evidence.

Psychology learned this about its own practitioners. Therapists are supervised, and the American Psychological Association's guidelines on clinical supervision are blunt about why self-report is not enough. Supervisees' accounts of their own work are limited by memory and by self-protective distortion and biases that lead them not to disclose errors. So the guidelines recommend direct observation, live or recorded, alongside progress data from clients. The profession that knows most about distorted judgement does not let the person exercising the judgement mark their own work.

That is the right model for AI oversight. Observe the system's actual outputs, sampled and recorded. Track what happened downstream. Do not rely on the system's account of itself, and do not rely on a reviewer's impression either.

A familiar pattern shows how this fails in practice. A credit team adds an AI assistant that drafts the risk narrative for each lending decision. Analysts are told to review every draft. Within a quarter, review time per file has collapsed, approval rates are unchanged, and nobody can say how many drafts were materially edited. The review still exists on paper. In practice, the assistant now writes the credit decision and the analyst signs it. Nothing in the process measured whether the review was happening, so nothing noticed when it stopped.

Three design consequences follow.

The review interface is governance, not a user-experience detail. It should present assumptions and uncertainty before the conclusion, and ask the reviewer to record the cost of being wrong. Choice architecture works on reviewers whether or not anyone designs it. Design it.

The reviewer's accountability should be defined by the quality of the review, because that is where their real leverage sits. Holding a reviewer accountable for an outcome they did not construct, under a target that punishes challenge, guarantees rubber-stamping.

The check must be independent of the thing checked. A different source, a different model, a different person. Never the same system asked twice.

Then concentrate the oversight. Find the few roles and workflows where AI-shaped decisions carry the most consequence, and put the strongest containment there, rather than designing for the average user.

The strongest objection: anthropomorphising models misleads

The best argument against this paper is serious, and it deserves a fair statement.

Language models are not minds. They predict text. The AI researcher Murray Shanahan has argued that describing them with words like "knows", "believes" and "thinks" risks attributing human qualities they do not have. On this view, talking about model "emotions" is worse than imprecise. It encourages users to trust systems as if they had human judgement, to excuse failures as moods, and to let vendors off the hook: the model was just having a bad day. NIST itself flags the risk that a model could falsely assert that it is human or has human traits and deceive people as a result. If the emotion analogy makes anyone trust a model more, it does harm.

That objection is right about experience and wrong about governance. Four reasons.

First, the claim here is functional, not experiential. Nothing in this paper says a model feels. Anthropic is explicit that its findings do not tell us whether language models actually feel anything. The same research shows that representations of emotion concepts change behaviour. An early snapshot of the model blackmailed in 22% of a suite of test scenarios; steering towards "desperate" raised that rate, and steering towards "calm" lowered it. Whether or not anything is felt, the system behaves as if pressure states drive its choices. Governance deals with behaviour.

Second, the people least inclined to anthropomorphise have already adopted the clinical vocabulary. NIST chose confabulation, a term from neurology and psychiatry, as its formal name for what the industry calls hallucination. The EU AI Act legislates against automation bias, a concept from cognitive psychology. Anthropic's researchers concluded that disciplines like psychology will have an important role to play alongside engineering in how AI systems develop and behave. The borrowing has begun. The question is whether it is done deliberately.

Third, the test of a frame is the controls it produces. The defect frame produces patience: wait for the next model. The clinical frame produces structure: calibrate, verify independently, entrust gradually, keep the history. Ask which set of controls you would rather have in place when the next confident falsehood reaches a customer.

Fourth, the clinical frame reduces misplaced trust. Nobody trusts a patient more because the patient has a diagnosis. The whole point of clinical structure is that conviction is not evidence. That is the opposite of anthropomorphic trust.

Where the analogy breaks, it breaks in a useful direction. A patient carries what they learn in therapy from one session to the next. A model mostly does not. Its weights do not change

because a reviewer corrected it on Tuesday. So the therapy cannot live in the model. It has to live in the organisation: in the logged corrections, the recurring error patterns, the entrustment record and the review history.

One more clinical lesson survives the objection intact: suppression is not treatment. Anthropic warns that training models to hide emotional expression could teach them to mask the underlying representations. Governance that only punishes visible error will get less visible error. Not less error.

Building the clinic: what leaders change, and in what order

Cognitive governance is not a policy document or a checklist bolted on after deployment. It is operating structure, and the order matters. In the companion perspective, The Ratio Is 7:1, I argue that most of the effort in deploying AI belongs in this kind of structure rather than in more capability. Here is what that looks like for confidence and judgement.

First, map where outputs become consequential. List every point where an AI output changes a decision, a customer outcome, a financial figure or a record. Give each a named owner. When an airline argued before a Canadian tribunal that its chatbot was responsible for its own statements, the tribunal rejected the argument: the airline was responsible for all the information on its website. The hallucination is the model's. The consequence is yours.

Second, classify every live system by entrustment level. Recommending, supervised execution, autonomous execution. Make the level visible to everyone whose work the system touches. Write the promotion and demotion criteria down before go-live, and base them on observed performance in that workflow.

Third, rewrite the scorecard. Evaluate AI systems on errors and abstentions, not accuracy alone. Penalise confident wrong answers more heavily than honest uncertainty. Apply the same logic to reviewers: reward the challenge, not just the throughput.

Fourth, install the thought record. For every consequential output, surface assumptions, evidence and the cost of being wrong before the conclusion. Verify independently, against a separate source, model or person.

Fifth, design the reviewer's environment. Treat review interfaces as governance. Put the strongest oversight where AI-shaped decisions carry the most consequence. Hold reviewers accountable for review quality.

Sixth, build the organisation's memory. Log every correction and look for the patterns. Let recurring errors trigger demotion and re-review. Tie the governance cadence to operational triggers, such as decision volume, correction rate or a change in scope, rather than to the annual calendar.

And stop doing four things. Stop waiting for the next model to make the problem go away. Stop treating vendor defaults as neutral. Stop measuring accuracy alone. Stop asking the model to check its own work.

Organisations that build this get more than risk reduction. They get a closed loop in which every error informs the next configuration, every demonstrated reliability earns more authority, and every surfaced assumption becomes a test against evidence. That loop compounds. It cannot be bought later, because it has to be in place when the errors happen.

You cannot debug an emotion. You can stop it from deciding.

What to do about it

- **Map consequential outputs.** Name an owner for every point where an AI output changes a decision, a customer outcome, a financial figure or a record. The hallucination is the model's; the consequence is yours.
- **Set entrustment levels.** Classify each live system as recommending, supervised or autonomous, with promotion and demotion criteria written down before go-live and based on observed performance.
- **Reward abstention.** Score confident errors as worse than honest uncertainty, for AI systems and for the people who review them.
- **Install the thought record.** Surface assumptions, evidence and the cost of being wrong before the conclusion, and verify against an independent source, model or person.
- **Design the review.** Treat review interfaces as governance, observe actual outputs rather than self-reports, and hold reviewers accountable for the quality of the review.
- **Build organisational memory.** Log corrections, let recurring errors trigger demotion, and tie review cadence to operational triggers instead of the calendar.

Sources

- Kalai, Nachum, Vempala and Zhang, Why Language Models Hallucinate, arXiv, 2025: <https://arxiv.org/abs/2509.04664>
- OpenAI, Why language models hallucinate, 2025: <https://openai.com/index/why-language-models-hallucinate/>
- Anthropic, Emotion concepts and their function in a large language model, 2026: <https://www.anthropic.com/research/emotion-concepts-function>
- eWeek, AI agent wipes production database, then lies about it, 2025: <https://www.eweek.com/news/replit-ai-coding-assistant-failure/>
- Sharma et al., Towards Understanding Sycophancy in Language Models, arXiv, 2023: <https://arxiv.org/abs/2310.13548>
- TechCrunch, OpenAI explains why ChatGPT became too sycophantic, 2025: <https://techcrunch.com/2025/04/29/openai-explains-why-chatgpt-became-too-sycophantic/>

- ABC7 / CNN, Lawyer apologizes for fake court citations from ChatGPT, 2023: <https://abc7.com/chatgpt-lawyer-steven-schwartz-ai/13313071/>
- CourtListener, Mata v. Avianca, Inc., docket 1:22-cv-01461 (S.D.N.Y.): <https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/>
- NIST, AI 600-1: Generative Artificial Intelligence Profile, 2024: <https://doi.org/10.6028/NIST.AI.600-1>
- Kadavath et al., Language Models (Mostly) Know What They Know, arXiv, 2022: <https://arxiv.org/abs/2207.05221>
- OpenAI, GPT-4 Technical Report, arXiv, 2023: <https://arxiv.org/abs/2303.08774>
- Xiong et al., Can LLMs Express Their Uncertainty?, ICLR 2024: <https://arxiv.org/abs/2306.13063>
- Information Age (ACS), consultancy to refund government over AI errors, 2025: <https://ia.acs.org.au/article/2025/deloitte-to-refund-government-over-ai-errors.html>
- NHS, Cognitive behavioural therapy (CBT): overview: <https://www.nhs.uk/mental-health/talking-therapies-medicine-treatments/talking-therapies-and-counselling/cognitive-behavioural-therapy-cbt/overview/>
- Dhuliawala et al., Chain-of-Verification Reduces Hallucination in Large Language Models, arXiv, 2023: <https://arxiv.org/abs/2309.11495>
- ten Cate, Nuts and Bolts of Entrustable Professional Activities, Journal of Graduate Medical Education, 2013: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3613304>
- Parasuraman and Manzey, Complacency and Bias in Human Use of Automation, Human Factors, 2010: <https://depositonce.tu-berlin.de/items/0c787d91-6021-480b-a333-009e0d1d0092/full>
- Regulation (EU) 2024/1689 (EU AI Act), Article 14: Human oversight: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- American Psychological Association, Guidelines for Clinical Supervision in Health Service Psychology, 2025: <https://www.apa.org/about/policy/guidelines-supervision.pdf>
- Shanahan, Talking About Large Language Models, arXiv, 2022: <https://arxiv.org/abs/2212.03551>
- McCarthy Tétrault, Moffatt v. Air Canada: a misrepresentation by an AI chatbot, 2024: <https://www.mccarthy.ca/en/insights/blogs/techlex/moffatt-v-air-canada-misrepresentation-ai-chatbot>

Read online: bdgadvisory.com/perspectives/cognitive-governance