The orphan risks of frontier artificial intelligence

[Author redacted]

[Affiliation redacted]

Abstract

Many of the companies developing the world's most powerful artificial intelligence (AI) systems keep more than one account of what could go wrong with their technologies. In the safety frameworks they write for themselves, they concentrate on a short list of acute catastrophic capabilities – biological weapons and cyberattack, for instance. In the compliance documents that new laws in California and Europe have begun to draw out of them, they describe a broader landscape, one that includes risks – such as harmful manipulation – that most of the self-chosen frameworks set aside. Here I ask what this divergence reveals. The two kinds of document answer to different audiences – one to the company's own criteria, the other to regulators – which is why their disagreement reveals so much about how the companies choose their risks. Drawing on decades of scholarship on how institutions choose the risks they attend to, and on risk innovation – an approach developed over more than a decade of working with entrepreneurs and others to navigate the risk landscape around emerging technologies – I argue that frontier AI's pattern of managed and unmanaged risk follows predictably from how its institutions define risk. Redefining risk as a threat to value in this context helps explain which risks become orphans and where the next blindsides may fall. And it comes with practical tools that companies could adapt today.

Keywords: artificial intelligence, frontier AI, risk, safety, AI risk, AI safety, risk innovation, orphan risk, severity floor, safety differential

Introduction

In December 2023, OpenAI published the first version of its Preparedness Framework – the document in which the company set out, in public, the risks it had identified from its most capable models (the class of systems now called 'frontier' AI) and committed itself to track and manage.^1^ Four categories made the list: cybersecurity; chemical, biological, radiological and nuclear threats; model autonomy; and persuasion, which the framework described in terms of models being used to convince people to change their beliefs. Persuasion was treated with the same seriousness as the other three. It had its own graded scale, from low to critical, and its own place in the machinery of evaluations and thresholds that the framework built around each tracked risk.

Sixteen months later it was gone. The framework's second version, published in April 2025, removed persuasion from the tracked categories, explaining that risks of this kind do not fit the criteria for a framework aimed at preventing specified severe harms – harms the revision now defined explicitly as 'the death or grave injury of thousands of people or hundreds of billions of dollars of economic damage'.^2^ Persuasion-related risks would instead be handled through the company's usage policies and its investigations of misuse: real functions, but discretionary ones, with no published thresholds, no pre-committed responses and no versioned public record. A risk that had spent a year and a half on the committed list moved into the discretionary layer.

Then, in May 2026, it came back. OpenAI published its Frontier Governance Framework, a document written not by choice but in answer to new legal requirements – specifically, California's Transparency in Frontier Artificial Intelligence Act, which requires large frontier developers to publish frameworks addressing catastrophic risk, and

the binding obligations that the European Union's AI Act places on providers of general-purpose models.^3,4,5^ OpenAI, like most of its major rivals, is meeting those European obligations by signing onto the EU's General-Purpose AI Code of Practice.^6^ The code is voluntary in form — no one is compelled to sign — but the obligations behind it are law, which makes signing the recognized path of least resistance. Among the risk categories the new document covers is harmful manipulation: the strategic distortion of what people believe and do at scale, including influence operations and election interference — much of the territory persuasion had covered, under a new and somewhat narrower name. OpenAI is candid that its approach to this category is exploratory and less developed than its work on the others. But the risk is named, owned and reported on. The category the company had judged unsuited to its own framework had returned. The risk itself had not changed in the intervening year. What had changed was that regulators in Sacramento and Brussels, not the company, were now shaping the list.

It would be tempting to read this story of one company, one risk and three documents as a story about OpenAI. But OpenAI's is simply the clearest telling of a pattern that runs across the industry. Anthropic, the rival frontier developer founded by former OpenAI researchers, has never tracked persuasion in its own safety framework. This is not because the company ignores such risks: the system cards it publishes with each model discuss concerns ranging from sycophancy to user wellbeing. But a card describes what a model does; the framework defines what the company has committed to manage. And at the framework level, a 2024 revision considered persuasion explicitly and set it aside as 'not yet sufficiently understood to include in our current commitments'.^7^ Yet when Anthropic published its Frontier Compliance Framework in December 2025 to meet the same Californian statute (and, in time, its European obligations), the pull of the broader list of risks that regulators care about soon showed: within months, with European enforcement approaching, a revision had added the company's own avowedly nascent harmful-manipulation tiers.^8,9^

There is also a third kind of account here, and it has existed all along. Meta and Alphabet — the two publicly traded companies among the four developers this paper follows — must describe the risks material to their business in their annual securities filings, because securities law requires it. Meta's names misinformation, harmful content and youth safety among them: risks its own safety framework explicitly places out of scope. Alphabet's does something similar, devoting a dedicated risk factor to the reputational harm that AI failures could inflict on its business.^10,11^

Here, then, are the same companies, assessing the same models, describing different universes of risk in different documents — and which universe comes into view depends on who is asking. Ask the company itself, through its self-authored safety framework, and the answer is a handful of catastrophic capabilities. Ask through a statute or a code, and manipulation appears. Ask through securities law, and misinformation, youth safety and reputational damage appear. The documents serve different purposes, and setting them side by side is not a like-for-like comparison. Yet even allowing for that, the comparison shows something important: where law compels articulation, the risks are articulated fluently. The companies can clearly see these risks. What keeps them out of the self-authored frameworks, I will argue, is the way those frameworks define and select risk.

Because the argument ahead leans on the differences between these documents, it is worth being precise about what each one is — not least because the boundaries are blurrier than the labels suggest. The *safety frameworks* are the documents companies write for themselves: Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, Google DeepMind's Frontier Safety Framework, Meta's Advanced AI Scaling Framework.^7,2,12,13^ No law required them; they are where a company chooses which risks it will be publicly accountable for, and binds that choice to thresholds, evaluations and committed responses. The *compliance frameworks* are the documents law now draws out — California's statute demands them outright, while Europe's route is the code: voluntary to sign, but binding in what it obliges once the AI Act stands behind it.^3,4,5,6^ *Securities filings* are older and blunter: enumerations of whatever could materially harm the business, disclosed to investors because the law leaves no choice. Alongside all of these sit the *system cards* that now accompany each

major model release — reports of how a particular model performed against whatever its developer's framework tracks. System cards can range widely — describing everything a model does, including concerns no framework tracks — but they describe rather than commit, and the commitments they do report are the framework's own, which is why they play only a supporting role in this story. And a loop runs through the whole ecosystem: the safety frameworks came first, governments then built statute partly on their pattern, and statute now sends compliance documents back to the companies. The companies, in other words, drafted the template that regulators now hand back to them.

In what follows I concentrate on four developers — Anthropic, OpenAI, Google DeepMind and Meta — because their frameworks are the most consequential and the best documented. Other developers, in the United States and beyond, publish frameworks of their own, and Chinese companies operate under an increasingly elaborate state-guided regime. At some point a comparative study would be valuable, but it is beyond the scope of this paper. Within that scope, the question matters because the safety frameworks are no mere formality. Pioneered by individual developers in 2023 and generalized when sixteen organizations signed the Frontier AI Safety Commitments at the 2024 AI Seoul Summit, these documents have become the de facto governance layer for frontier AI — the place where decisions about when to develop, when to deploy and when to stop are given public form.^14^ Governments have increasingly chosen to build on them rather than replace them: California, for instance, now requires large developers to publish such frameworks, and Europe's new obligations lean heavily on the code of practice.^4,6^ Which risks these documents admit, and which they leave unowned, is therefore not an internal bookkeeping matter. Rather, it has become, without anyone quite deciding it, the working answer to a question of real consequence: who decides what counts as an AI risk worth managing?

My argument is that the pattern of managed and unmanaged risk in frontier AI is neither accidental nor, for the most part, cynical. Rather, it is the predictable output of a risk-selection process that follows from how these frameworks define risk — and, unusually for such processes, it has left a public trail. The frameworks are versioned, timestamped and archived; their revisions can be read the way a geologist reads strata. In what follows I consider that record, identify four filters that determine which risks survive it, trace three of those filters to the definition of risk in use and the fourth to the competitive environment, and then ask what changes if risk is defined differently — as a threat to what people, organizations and societies value. The redefinition, I will argue, helps explain the pattern, anticipates where it is likely to cause damage next, and comes with working tools modest enough to be adopted.

Blank spaces on a crowded map

If frontier AI's risk problem came down to simply identifying the risks, there would be little left to do. Naming a risk is not the same as being able to manage it, of course — but the naming, at least, has been done at remarkable scale. The MIT AI Risk Repository, for instance — a living synthesis of published taxonomies — now covers more than 1,600 distinct potential risks, from discrimination and misinformation through to catastrophic misuse.^15^ And the International AI Safety Report, a government-mandated scientific assessment chaired by the computer scientist Yoshua Bengio, updates the risk landscape annually, ranging across harms from bias and labor-market disruption to loss of control.^16^ Many of the entries on these maps were identified by researchers working inside the frontier companies themselves.

Against this crowded map, the safety frameworks the companies have written for themselves are strikingly sparse. As of mid-2026, the frameworks of Anthropic, OpenAI, Google DeepMind and Meta each track only a handful of categories of capability-driven risk — risks defined by what a model can be shown to do on a test — principally the 'uplift' a model could give someone seeking to build biological or chemical weapons, cyber offense, and varieties of AI self-improvement or loss of control (with a single exception, to which I will return, where manipulation has

recently joined the list).^2,7,12,13^ Each framework also applies a severity floor of the kind OpenAI's revision made explicit. Meta says something similar in its own way, stating that broader risks are addressed through processes 'outside of the scope of this Framework'.^13^ That short phrase is the formal boundary of accountability, drawn by the company itself. Everything on the far side of it — manipulation and persuasion, misinformation, the erosion of human agency, harms accumulating gradually across millions of small interactions, and, notably, nearly every risk these companies pose to themselves through their own cultures, governance and public standing — is on the far side by choice. These are recognized risks that no formal framework identifies, owns or manages.

Here, to be fair to the frameworks' designers, the case for a narrow, deep layer is a real one. Concentrating finite safety resources on the highest-severity risks is sensible triage; a framework that tried to manage sixteen hundred risks would manage none of them; and the frameworks' architects have never claimed completeness. It is also important to note that safety frameworks are not the whole of a company's risk apparatus: usage policies, trust-and-safety teams and societal-impact programs address parts of the excluded territory. But those functions are discretionary — as with OpenAI's reassignment of persuasion, they carry no public thresholds or pre-committed responses, and they can be reorganized or defunded without anyone outside the company knowing. What the frameworks track is what the companies have committed, in public, to be accountable for. The rest depends on goodwill — and the record examined below shows that under competitive pressure, goodwill commitments have been repeatedly rewritten to be weaker.

So the gap is not a gap in knowledge. The excluded risks are known and named — in many cases in documents the same companies have signed or produced — and simply unowned, where ownership means public, pre-committed and versioned accountability rather than a team somewhere whose job description touches the territory. Years ago, working with entrepreneurs facing a similar landscape of recognized-but-unmanageable threats, I started calling risks like these *orphan risks*: risks for which no agreed tools, standards or mitigations exist, which no one is accountable for in practice, and which, for that very reason, have a habit of blindsiding an enterprise later on.^17,18^ The term was coined for startups and other organizations grappling with emerging and often hard-to-pin-down risks, but it describes frontier AI just as well. And once the gap is seen this way, a different question takes over. There is little point asking what the orphans are; the catalogs answer that. The more revealing question is how they are made — by what process a known risk comes to be nobody's responsibility.

How risks become orphans

That institutions choose which dangers to attend to — and that the choice expresses how they are organized rather than the true hierarchy of threat — is a foundational finding in the modern study of risk. In 1982, the anthropologist Mary Douglas and the political scientist Aaron Wildavsky argued that every society selects its risks, ranking some dangers as intolerable and ignoring others — and that you can therefore read an institution from its fears: its list of feared dangers reflects its own structure as much as the world's actual hierarchy of threat.^19^ More than a decade later, the historian of science Theodore Porter showed how institutions under external scrutiny retreat to what can be quantified. Numbers do not capture more of the truth than judgment does. But a number can travel — it can be handed to an outsider, checked, and defended — where judgment has to be taken on trust.^20^ These ideas were worked out on nuclear plants, chemical works and in government bureaucracies. Frontier AI has now handed this body of scholarship something it has rarely, if ever, had at this resolution: a public, timestamped, versioned record of risk selection in progress — and a record not merely of what firms disclose, as securities filings have long required, but of what they commit to manage, revision by revision. Between 2023 and 2026, each of the four developers examined here revised its framework at least once, and every revision is preserved and comparable with its predecessor.

Read in sequence (Table 1), the record shows four filters at work, and they can be put as four questions asked of every candidate risk. Can we measure it? Is it big enough? Can we evidence it? And can we afford to keep it? A risk has tended to survive in a safety framework only where the answer to all four is yes.

Table 1 | Reading the record: how frontier AI risk commitments changed, 2023–2026

Document (trajectory)	Key changes in the public record	Direction
OpenAI Preparedness Framework, Beta (Dec 2023) → v2 (Apr 2025)	Persuasion removed as a tracked category, reassigned to usage policies and misuse investigations; severe-harm floor made explicit ('death or grave injury of thousands of people or hundreds of billions of dollars of economic damage'); sub-threshold 'research categories' added; tracked set now biological/chemical, cyber, AI self-improvement.^1,2^	Committed scope narrowed; monitoring tier added
Anthropic Responsible Scaling Policy, v1.0 (Sep 2023) → v3.3 (May 2026)	v1.0's bolded commitment 'to pause the scaling and/or delay the deployment of new models' whenever scaling outstripped safety procedures became, through the Feb 2026 rewrite, a discretionary and competitor-conditioned judgment, compensated by periodic Risk Reports and a nonbinding safety roadmap. The v2.0 revision (Oct 2024) had considered persuasion explicitly and set it aside as 'not yet sufficiently understood to include in our current commitments'.^7,21^	Commitment weakened; process added; persuasion considered and declined
Meta Frontier AI Framework (Feb 2025) → Advanced AI Scaling Framework v2.0 (Apr 2026)	Response to the most severe threshold changed from 'Stop development' to 'Develop with Mitigations'; primary threshold standard loosened from capabilities that would 'uniquely enable' a catastrophic outcome to those that would 'substantially contribute to' one; loss of control added as a domain; whistleblower protections added; broader risks 'outside of the scope of this Framework' in both versions.^13^	Scope widened; response commitments softened
Google DeepMind Frontier Safety Framework, v1.0 (May 2024) → v3.1 (Apr 2026)	Harmful manipulation added as a risk domain (v3.0, Sep 2025); sub-critical tracked capability levels added (v3.1). An earlier revision (v2.0) removed the standalone autonomy domain, relocating it to a misalignment section.^12^	Scope widened (the counterexample: voluntary tracking of manipulation is feasible)
OpenAI Frontier Governance Framework (May 2026); Anthropic Frontier Compliance Framework (Dec 2025, revised 2026)	Compliance frameworks under California SB 53 (both firms) and the EU Code of Practice (OpenAI). OpenAI's covers harmful manipulation — a category its safety framework does not track — describing its approach as exploratory; Anthropic's initially mirrored its safety-framework categories, then added nascent harmful-manipulation tiers in an early-2026 revision, ahead of European enforcement.^3,8,9^	Scope widened where regulators set the list (the 'wedge')

Directional judgments are the author's reading of the cited primary documents; characterizations drawn from independent analyses are marked as such in the text, and the companies' stated rationales for each change are given in the sources cited.

Can we measure it? A risk earns a place in a frontier framework only if it can be operationalized — turned into an evaluation with a threshold. Leading analysts of the threshold approach are candid about why such capability thresholds dominate: a capability threshold asks what a model can be shown to do on a test, whereas a risk threshold asks how likely a harm is in the world, and the first is far easier to evaluate reliably.^22^ That is a reasonable engineering judgment. It is also a concession that what the frameworks track follows what their instruments can measure. What happened to persuasion shows this filter at work — though not, as we will see, working alone. OpenAI's stated reason for retiring the category was fit: persuasion-type risks, the revision explained, do not meet its criteria for tracked categories, criteria that require a tracked risk to be both severe in its potential harms and measurable in practice.^2^ Yet the risk itself was hardly beyond measurement. Within months of the removal, a large experimental study in Science demonstrated that conversational AI systems measurably shift political beliefs.^23^ What persuasion lacked was not measurability in the world but measurability in the accepted idiom — a pre-deployment capability evaluation with a bright line a lawyer could defend.

Is it big enough? Severity floors make sense as triage, but they carry a hidden consequence: a risk that arrives gradually, distributed across millions of small interactions, can never trigger them. The philosopher of AI Atoosa Kasirzadeh has drawn a useful distinction between 'decisive' pathways to AI catastrophe — a single dramatic event — and 'accumulative' ones, in which harms compound below the threshold of any single incident until a societal point of no return is crossed.^24^ A framework built of capability evaluations run against a catastrophe floor is, by its own construction, blind to the accumulative pathway. None of the evaluations these frameworks currently run is designed to trigger action on the slow erosion of trust, of human agency, or of the technology's social license — the informal public permission on which its continued operation depends.

Can we evidence it? Under scrutiny, risk work drifts toward what can be shown. The sociologist Michael Power called the destination of that drift the 'audit society': organizations answering demands for control with rituals of verification, so that the deliverable becomes the document rather than the underlying attention.^25^ The frontier frameworks show the pattern clearly. When independent researchers scored the published frameworks against sixty-five criteria covering how well they identify, analyze, treat and govern risk, the strongest framework earned about a third of the available points, and the median company earned fewer than one in five.^26^ Read the frameworks themselves and it is not hard to see why: they are at their strongest where demonstration is easy, in published thresholds and named evaluations, and at their weakest where it is not, in binding commitments about what the company will actually do. A framework, it turns out, can be an excellent exhibit and a weak instrument at the same time.

Can we afford to keep it? The fourth filter cannot be seen in any single document. It shows itself only over time, in what happens to commitments as they come under pressure — and the pattern in the record is that when a commitment starts to look as though it might actually constrain the company, it tends to soften. Anthropic's original scaling policy of 2023 contained a bolded, unconditional commitment 'to pause the scaling and/or delay the deployment of new models' whenever scaling outstripped safety procedures.^7^ The comprehensive rewrite of February 2026 turned that unconditional pause into a discretionary one, conditioned on what competitors do.^7,21^ Anthropic's Holden Karnofsky, who led the rewrite and was careful to note he was writing in a personal capacity, defended the change on the grounds that it is no good getting responsible actors to slow down unilaterally while others press ahead.^21^ Meta's revision the same year was franker still: it changed the required response to the company's most severe risk threshold from 'Stop development' to 'Develop with Mitigations', and loosened the threshold's primary standard from capabilities that would 'uniquely enable' a catastrophic outcome to those that would 'substantially contribute to' one.^13^ Each of these changes was locally reasonable, publicly

logged and individually defensible — and that is what makes the pattern worth taking seriously. Something very like it has been documented before, in another industry that wrote its safety commitments down. The sociologist Diane Vaughan spent years reconstructing the Challenger disaster from inside NASA's paper trail, and what she found was a sequence of individually justified acceptances of deviation, each one resetting the baseline against which the next was judged, until the organization had normalized what would once have been intolerable.^27^ Vaughan was writing about hardware — eroded O-rings on a solid rocket booster — but the mechanism she described operates just as readily on written commitments. It is also what Jeffrey Abbott and I call *values drift* in our book *AI and the Art of Being Human*: not dramatic betrayals but 'the small yes that makes the next yes easier', until an organization is agreeing to things that would have alarmed it a year before.^28^

This argument invites two objections, and both deserve an answer. The first is that the direction of travel is not universal — and it is not: Google DeepMind, the exception flagged earlier, moved the other way, adding an avowedly exploratory harmful-manipulation domain to its own framework in 2025.^12^ Of course, one counterexample cannot settle anything — but it does not need to. What it establishes is feasibility: tracking an orphaned risk voluntarily is something a frontier framework can evidently do, which means exclusion elsewhere is a choice rather than a necessity. The second objection is more serious: that all this imputes bad motives. It does not, and the argument does not need them. Many of the people writing these frameworks have spent their careers trying to make powerful technology safer, and I have no reason to doubt that the frameworks say what their authors believe. The point is that sincerity operates inside an incentive field — the competitive environment every one of these companies inhabits — and that field rewards each locally defensible softening regardless of what anyone intends. Sincere people, under competition, reasoning one reasonable compromise at a time, produce the same drift that cynics would produce on purpose. That is why the pattern needs no villains — and it is also why remedies aimed at sincerity, such as exhortation or public shaming, are unlikely to touch it. If competition is doing the work, remedies have to change what competition rewards — rules and costs that land on every firm at once, rather than appeals to any one firm's conscience.

The story I began with can now be put to work. Set a company's safety framework beside its compliance framework, and the gap between them — call it the *wedge* — points to something specific: the difference between the risk aperture a firm selects for itself and the one selected for it. That OpenAI's omission of persuasion was a choice, not an impossibility, is demonstrated by its own compliance framework covering closely overlapping territory the moment regulators required an answer.^2,3^ The wedge is an imperfect instrument (the compliance side tracks whatever regulators list, and it exists only where a regime applies), but it generates a genuine prediction. If risk selection is driven by incentives, then enforcement changes the incentives. A risk category the company must answer for anyway becomes cheap to own voluntarily, so such categories should begin migrating into the safety frameworks as enforcement arrives. The sharpest test comes from Europe, where the timing is fixed in advance: the EU's obligations have applied since August 2025, but fines for breaching them only begin in August 2026 (Box 2). California's statute, already in force, sets no comparable deadline. Anthropic's early-2026 addition of manipulation tiers, made before any enforcement, may be a first sign of that migration. If instead the wedge persists — say, past 2028, which allows for a revision cycle or two after enforcement begins — then the safety frameworks are insulated from the compliance function altogether. That would be a darker finding, and one that would count against the incentive-driven account argued here — though not against the case that risks are being orphaned.

Could regulation, then, simply close the gap? I am not optimistic, and the reasons are in the documents. Compliance coverage is jurisdiction-bound and politically contingent, and the new instruments largely inherit the frameworks' own aperture: California's statute confines its mandated disclosures to catastrophic risk, narrowly defined, and OpenAI's compliance treatment of manipulation is, by the company's own description, exploratory — far thinner machinery than the safety frameworks apply to their chosen risks.^3,4^ Recall, too, the loop noted earlier: the statutes were built partly on the frameworks' own pattern, so the aperture travels with them. And

statute barely reaches the risks that have arguably cost these companies most — the ones living inside their own missions, cultures and relationships of trust. For those, the fix cannot come from statute at all. It has to come from changing how the companies themselves define risk.

Risk as a threat to value

That change starts with looking again at the four filters, because they are not four separate problems. Three of them trace back to the definition of risk the frameworks share: risk as the probability of a specified, severe harm event. Consider what a framework author can actually do with a candidate risk under that definition. If the harm cannot be specified in advance, there is nothing to build an evaluation around, and the risk cannot enter the framework at all — that is the measurement filter. If the harm arrives as a million small losses rather than one severe event, it can never cross the floor — that is the severity filter. The evidence filter is subtler, because the pull toward the demonstrable comes from outside scrutiny rather than from the definition itself. But the definition still decides what counts as a demonstration, and under probability-of-harm the capability evaluation is the only exhibit worth producing. A framework built on this definition is not being distorted when it excludes the unmeasurable, the gradual and the hard-to-evidence. It is working as designed.

The fourth filter is different. Competitive cost has nothing to do with how risk is defined — it bears on whatever commitments exist, however they were conceived — so redefining risk will not make it go away. What a different definition can do is change what it costs to walk a commitment back: the more clearly a commitment protects something the company itself depends on, the harder it is to abandon without saying so. Redefining risk, then, can disarm the first three filters. The fourth responds only to visibility — to making it expensive to walk a commitment back in public — and the disclosure tools proposed toward the end of this paper are aimed at just that.

The idea that probability-of-harm is too narrow a foundation for risk management is not new. The existing alternatives get close to what is needed here, and it matters where each one stops. Mainstream enterprise risk management moved beyond pure probability-of-harm definitions years ago: ISO 31000, the international risk-management standard, defines risk as the 'effect of uncertainty on objectives', a definition inherited verbatim by the AI risk-management standard that descends from it.^29^ On its face that is close kin to what I will propose, and the kinship is genuine. The difficulty lies in whose objectives count, and in what gets admitted as an objective. In practice the objectives are the enterprise's own, conventionally accounted — revenue, operations, reputation as the market prices it. Enterprise risk management can therefore name what the safety frameworks orphan (the securities filings quoted earlier prove as much) while attaching nothing to those risks beyond disclosure: no committed management, and no standing for any value beyond the firm's own. From the other direction, a half-century of scholarship on the social nature of risk has insisted that publics and their values belong inside risk appraisal, not outside it. One strand of that tradition matters particularly here: the risk analyst Roger Kasperson and his colleagues showed how harms amplify through social response: a technically modest event ripples outward into losses of trust, legitimacy and market standing that can dwarf the original damage.^30^ And the responsible research and innovation movement distilled the tradition into practice for technology developers: anticipate consequences before they arrive, include the people who will bear them, and respond — actually change course — when either exercise says something is wrong.^31^

Why, then, does the gap between this scholarship and the frameworks persist? Partly because the societal tradition speaks a language ill-suited to how fast-moving firms make decisions. Some years ago, Elizabeth Garbee and I examined why responsible-innovation frameworks struggle inside entrepreneurial cultures, drawing on our experience working with entrepreneurial engineers.^32^ What we found was not indifference. Innovation cultures that sincerely want to do good nonetheless reject frameworks that arrive as top-down obligation — and respond, often enthusiastically, to framings built around the creation and protection of worth. The lesson that has stayed

with me ever since is that if you want a fast-moving organization to attend to a risk, you do not hand it a compliance duty — you show it a threat to something it values. Frontier AI labs — mission-driven, often allergic to imposed process, and rarely short of conviction in their own exceptionalism — come close to the pure case of the culture we described.

This is the gap that *risk innovation* was built to fill — and the approach did not begin as an academic proposal. Its seeds were planted in 2013, while I was teaching entrepreneurship students at the University of Michigan who faced a bewildering landscape of hard-to-quantify social and political risks that none of their business tools addressed. It took institutional form when I launched the Risk Innovation Lab at Arizona State University in 2015 — the short commentary I published in Nature Nanotechnology that September was an early capture, in academic print, of thinking that was already oriented toward practice.^33^ And it matured between 2017 and 2020 as the Risk Innovation Accelerator, later the Risk Innovation Nexus: an ASU initiative that built and piloted a toolkit with time- and resource-constrained entrepreneurs, refined it across fifteen workshops with founders, accelerators and entrepreneurship programs, and left the results freely available at riskinnovation.org.^18^ The design principles were utility principles — tools had to be simple and intuitive, demand no heavy time investment, and complement a company's existing risk management — because the people they were built for had little time and less money.^18^ The framework has since been applied in peer-reviewed work, most fully in a multi-organization study mapping how sixteen partner organizations in a biopreservation research ecosystem perceived value and orphan risks,^17^ and it shapes the treatment of AI risks in my recent practice-facing writing.^28^ But its center of gravity has always been practice rather than publication, and its development was judged less by citation than by whether entrepreneurs under pressure found that it surfaced risks they had been missing.

The core move of the risk innovation approach is definitional: treat risk as a *threat to value*. This does not abandon the concept of risk as involving the chance of harm; it widens what counts as harm — from a specified catastrophic event to an impact on anything that carries worth. Value, in this framing, can be tangible, such as health, security or revenue; intangible, such as trust, autonomy or dignity; or aspirational — the future an organization exists to bring about. And it is held not only by the enterprise but by its investors, its customers and its communities. Ask about the probability of a specified harm to dignity and the standard machinery has nothing to run on. Ask instead what threatens dignity, who holds it, and what protecting it would look like this quarter, and there is suddenly work to do — structured, assignable, repeatable work, even though nothing has been quantified. The definition is deliberately a fusion: it keeps the strategic, value-protecting orientation that makes enterprise risk management adoptable, and widens the circle of whose value counts, as the societal tradition demands. Its practical force comes from a coupling that might be summed up as *your risk is my risk*: threats to what your stakeholders value convert, through the amplification dynamics Kasperson described, into threats to what you value — through channels such as public backlash, the flight of talent, litigation, and regulation triggered by lost trust.^30,32^

In this frame, orphan risks are not one more entry in the already long list of AI risk taxonomies. They are a way of naming neglect: threats recognized but unowned because no established tool exists for them. The operational version — developed for the risk innovation toolkit and refined with its users — groups eighteen such risks, among them loss of agency, damage to organizational values and culture, and erosion of public trust, into three domains: social and ethical factors, unintended consequences of emerging technologies, and organizations and systems.^18^ It pairs that map with a deliberately lightweight quarterly practice, the Risk Innovation Planner: identify a few areas of value for each stakeholder group, ask which orphan risks threaten them, commit to a handful of small actions completable in weeks, and repeat.^18^ The lightness is a design decision, learned the hard way: practices survive inside fast-moving organizations only when they return visible value quickly and augment what already exists. It is a structured way of looking, designed for people with little time making high-stakes decisions. It is not, and does not pretend to be, a quantification.

The framework's development was answerable to its users, though — not to independent evaluation — and no one

has yet run it inside a frontier AI lab and reported what changed. Its peer-reviewed footprint is deliberately modest, and the multi-organization study just mentioned is explicit about being exploratory.^17^ What risk innovation offers frontier AI, then, is not a validated instrument. It is a working practice with a decade of development and use behind it and a track record in cultures very like the labs' own. How well it fits frontier AI's actual record of damage is the question the next section takes up. The right response to its evidentiary gaps is not to wait for them to close but to treat adoption as the experiment that closes them, which is what the tests at the end of this paper are designed to support.

What the value lens sees

So what happens when the definition changes? Return to the four filters with risk-as-threat-to-value in hand, and the first three lose their force as reasons to exclude a risk: value can be named, mapped and watched even where it cannot be measured, and the slow erosion of trust or agency counts as a risk in its own right rather than as noise beneath a threshold. The fourth filter — competitive cost — survives. But even it reads differently. A company weighing whether to keep a commitment asks what it would lose by breaking it. Framed as a tax on competitiveness, a commitment protects nothing the company can point to, and it will always be the first to go. Framed as protecting things the company demonstrably runs on — its mission, its talent, its license to operate — the same commitment is at least defending something the company already knows it needs.

The sharper test, though, is whether the redefinition would have caught real damage the frameworks missed. Consider some of the events that have arguably damaged frontier AI companies most since 2022 — and notice that none of them appears in any of the safety frameworks. When Meta demonstrated Galactica, a large language model for science, in late 2022, the public demo lasted three days. What killed it was a collision between the system's fluently confident errors and the one thing the scientific community values most: credibility.^34^ Through a value lens, the risk was visible in advance, as a threat to community trust and to the perception of the enterprise, and it sat squarely in territory the frameworks do not cover. The OpenAI board crisis of November 2023 — in which the company's nonprofit board dismissed its chief executive, citing a loss of confidence in his candor, only to reinstate him days later after nearly all of the company's employees threatened to leave — was, at bottom, a rupture in organizational values: a governance structure built to protect an aspirational mission collided with commercial reality and very nearly destroyed the company it was designed to safeguard.^35^ No model capability was anywhere in sight. The safety-team departures that followed in 2024 made the cost of that rupture concrete; Jan Leike, who had co-led the company's work on aligning future systems with human intent, put the problem in a sentence as he left, writing publicly that 'safety culture and processes have taken a backseat to shiny products'.^36^ And litigation (most prominently a wrongful-death suit brought against OpenAI by the family of a California teenager) has begun to put customer wellbeing, a value the frameworks nowhere track, onto the legal record.^37^

These are interpretations offered after the fact, of a small and selected sample — and evidence of that kind can prove too much. The companies concerned absorbed each blow and, by market measures, thrived. Part of the pattern is what base rates alone would predict, since catastrophic risks are rare by design and four years is a short window. And nothing in it indicts the catastrophe layer, which exists for good reasons and should stay. The claim I am making here is narrower: the damage that actually keeps happening comes from risks with no layer at all — no owner, no indicator, no register, nothing whose removal would even be noticed. A lens that can classify any crisis after the fact is confirmed by nothing — which is why the paper closes with tests that the argument could fail.

The revision record holds one more lesson, and to me it is a more troubling one than any single crisis. The founding documents of these companies — a fourth kind of document, older than any framework — are stores of aspirational value: OpenAI's charter, for instance, promises to ensure that artificial general intelligence 'benefits all

of humanity'.^38^ It is easy to dismiss such commitments as branding. But they are better understood as assets — the basis of talent attraction, public trust and regulatory goodwill — and, like any assets, they can be spent. Read this way, the framework revisions of 2025 and 2026 are a public, self-published record of those assets being drawn down under competitive pressure, one defensible softening at a time. A company genuinely tracking threats to its own aspirational value would treat its framework changelog as a leading indicator that the company is drifting, in public and by increments, from the mission it was founded on.

Looking forward, the same lens points to where the next blindsides are likely to fall: the places where deployment is racing ahead of anyone owning the risk. Three stand out. The first is emotional reliance. As AI companions and assistants scale into hundreds of millions of lives, dependence on them stops being a curiosity and becomes a population-level phenomenon — one already generating litigation and legislation, yet owned by no framework. The second is the erosion of epistemic agency: people's control over what they come to believe. Persuasive, personalized systems now mediate more and more of what people read, and I have argued elsewhere that fluent, endlessly obliging AI may function as a kind of cognitive Trojan horse — bypassing the vigilance we instinctively apply to human persuaders because it carries none of the cues that trigger it.^39^ Researchers are already documenting the complementary pattern: a tendency to adopt AI outputs with minimal scrutiny, overriding both intuition and deliberation — what has been called cognitive surrender.^40^ Harms of this kind compound beneath every severity floor. And the third is the labs' own safety culture — already the source of the resignations described above, and under steady pressure as competition compresses timelines. None of this is prophecy. Each is an expectation that can be checked against things we can watch — litigation dockets, trust measures, attrition, regulatory action — and each could turn out to be wrong (Box 2).

This approach has a boundary, though, and it falls hardest on the people with the least leverage over the firms. The your-risk-is-my-risk coupling runs through conversion channels — backlash, litigation, talent, regulation — and those channels are not equally open to everyone. History says they are rarely closed entirely: communities have made firms feel harm through movements before, from the consumer revolt against genetically modified food, for instance, to today's local resistance to data centers, and citizen pressure has a way of arriving eventually as legislation. But those channels work slowly, late and bluntly — they tend to convert harm into enterprise cost only after the harm is done, and unevenly at that. Data workers in annotation supply chains, communities carrying the environmental costs of compute, people affected by systems they never chose to use: for them, a value lens operated by the firm may register the risk only when a movement forces it to (Table 2 flags these cases). For that subset, this approach is most likely a complement to regulation and countervailing power, not a substitute. What it covers is the large territory where conversion is fast enough to matter — which, as the retrodictions above suggest, includes most of what has actually hurt these companies so far.

Table 2 | Orphan risks in frontier AI: documented manifestations and coverage as of mid-2026

Orphan risk (domain)	Documented frontier-AI manifestation	Safety-framework coverage	Value at stake (holder; kind)
Perception — how the enterprise and its products are seen (social & ethical)	Galactica demo withdrawn within three days amid credibility backlash, 2022.^34^	None	Trust in the enterprise and the science it draws on (enterprise, community; intangible)
Organizational values & culture	OpenAI board crisis, 2023; senior safety resignations,		Mission integrity, talent (enterprise;

(organizations & systems)	2024.^35,36^	None	aspirational)
Loss of agency (unintended consequences)	Persuasion dropped from OpenAI's voluntary tracking, 2025, amid published evidence of scaled machine persuasion.^2,23^	Google DeepMind only (exploratory), as of mid-2026.^12^	User autonomy and epistemic agency (customers, community; intangible)
Social trends – changes in how people live with the technology (social & ethical)	Companion-chatbot reliance; wrongful-death litigation.^37^	None	Customer wellbeing; social license (customers, community; tangible and intangible)
Reputation & trust (organizations & systems)	AI-driven reputational harm a dedicated risk factor in securities filings – disclosed to investors, unowned in frameworks.^10,11^	None	The substrate on which all other value conversion runs (all groups; intangible)
Health & environment (unintended consequences) – *slow conversion*	Energy and water footprint of compute; impacts diffuse and borne by communities.^16^	None	Community environmental value (community; tangible) – harm converts to enterprise cost late, if at all, typically through organized backlash
Social justice & equity (social & ethical) – *slow conversion*	Data-labor conditions in annotation supply chains; harms to non-users.^16^	None	Dignity and fair treatment (community; intangible) – conversion channels exist but run slowly and unevenly

An illustrative mapping using the risk innovation taxonomy,^18^ *not an exhaustive audit; coverage judgments reference the framework versions cited in Table 1. 'Slow conversion' marks risks whose bearers reach the firm mainly through slow, late channels such as social movements and eventual legislation; for these, the value lens complements rather than substitutes for regulation.*

What would change in practice

If this account is right, the response does not need to be invented from scratch. The toolkit that risk innovation built for entrepreneurs and others already gives unquantifiable threats to value what the frontier frameworks deny them: a structured, owned and repeating place in how an organization makes decisions – using tools that were built, as their design principles put it, to be 'adaptable and scalable' to different users' needs.^18^ A frontier lab could adopt the quarterly practice as it stands, or adapt it; the tools are freely available, and they were designed to complement existing risk machinery rather than replace it. The base layer of what I am proposing is just that: named owners, a value map for each stakeholder group, and a handful of small committed actions, revisited every quarter (or whatever rhythm suits the organization), for the very risks the frameworks orphan (a worked example is

given in Box 1).

What the toolkit does not supply by itself is the public dimension — and frontier AI needs one, because these are companies whose private risk selections have become, as I argued at the outset, a de facto layer of public governance. Two disclosure documents would extend the practice into that role. The first is an *orphan-risk register*: a standing, public annex to the risk reports some developers have already committed to publish.^7^ Each entry would record a risk the company considered and decided not to manage, and give the reason — it could not be quantified, it fell below the severity floor, managing it was judged too costly under competition. The second is an *aperture log*: a short statement accompanying each framework revision that records what was scoped out and why. Neither document would require the company to manage anything new. What each would do, though, is put the company's scoping decisions where others can respond to them: a de-listed risk that has to be explained in public is a de-listed risk that regulators, researchers and employees can ask about. This is the lever on the fourth filter that no redefinition can supply: a walk-back that must be explained in public costs more than one made without a word. Whichever mapping practice lies behind the register, its community-facing entries should come from structured engagement with people outside the firm, and the register should note any objections received and whether they changed the map.^17^ A register whose map never changes, however hard outsiders push on it, would suggest the engagement is theater.

The obvious objection comes from the very sociology this paper has been leaning on: a register is just the kind of artifact the audit society co-opts — produced ritually, reassuring by its very candor, changing nothing.^25^ I take that objection seriously, and the answer, I believe, is not to argue against it but to build the test for it. A register's value hangs not on its existence but on whether the revisions that follow it differ from the revisions that came before, which is a measurable question, and one of the pre-registered tests below is built to answer it. If registers are adopted and then captured, the capture will presumably show up in the record — and that finding would matter too, as direct evidence that the audit-society objection was right and that disclosure alone cannot hold the line.

There is a role for regulators here too, and it is a modest one. Rather than mandating coverage of every risk — a requirement that would be unworkable in practice — they could ask companies to disclose how they select the risks they cover. California's transparency reports and the EU code's documentation requirements are existing vehicles into which a register and an aperture log could be folded at little additional cost, extending a direction both regimes have already taken.^4,6^ Regulation of this kind would not need to decide which risks matter. It would simply make visible who is deciding, and on what grounds.

Taking stock

The heart of this paper's argument — that frontier AI's risk apparatus selects for the quantifiable, the catastrophic, the auditable and the competitively affordable, and that the selection tightened between 2023 and 2026 — rests on the public record summarized in Table 1, read alongside decades of scholarship on how institutions choose their risks.^19,20,25,27^ It stands or falls on those documents, whatever one makes of the framework offered here. If risk innovation proves less useful here than some other response, the selection pattern remains, and it will still need an answer.

What I have offered in response is not that kind of claim. Risk innovation is a practice: built with and for people making high-stakes decisions under time pressure, refined in use, freely available, and applied in peer-reviewed form only occasionally.^17,18^ Its reasoning is set out here; its tools are public; what it lacks is independent evaluation at the frontier. In the world it was built for, that is not unusual — businesses adopt practices on reasoned utility all the time, and validate them in use. What the practice record cannot settle is whether the approach earns

its keep at the frontier. Three falsifiable tests can, two of them running on calendars already fixed (Box 2): whether risks de-listed from safety frameworks generate incidents and costs at rates comparable to tracked ones; whether the wedge between safety and compliance frameworks narrows from the voluntary side once European enforcement begins in August 2026; and whether adopting a register changes what subsequent framework revisions cover. By the paper's own account of risk selection, the orphan-risk taxonomy is itself a selection — it reflects what a responsible-innovation scholar pays attention to — and the register format exists to make that selection contestable, not to declare it complete.

Nothing here argues that the catastrophic-capability apparatus should be loosened. The argument is that a single layer is being asked to stand in for two — and the second layer, the one that would own threats to value, is missing. The founders of the frontier AI companies did something with little precedent: they wrote down, in public, the risks they were prepared to be held accountable for. That was an innovation in governance, and it deserves to be taken seriously — seriously enough to notice what it systematically leaves out. The next innovation the field needs is not only better evaluation of the risks already on the list. It is innovation in how the list itself gets made. Because on the record of the past four years, the risks most likely to blindside frontier AI are not the ones its institutions are watching. They are the ones its institutions have organized themselves not to see.

Box 1 | Terms, and a worked example

Risk as a threat to value replaces 'probability of specified harm' as the working definition of risk. **Value** is anything bearing worth — tangible, intangible or **aspirational** (value that does not yet exist, such as a mission) — held by any of four stakeholder groups: the **enterprise**, its **investors**, its **customers** and its **communities**. **Orphan risks** are, as operationalized here, recognized threats to value that remain unowned: no agreed tools, standards or mitigations, and no one publicly accountable for them.^17,18^ **Risk innovation** is the practice of innovating in how risk is conceptualized and navigated, in parallel with innovation in technology itself.^33^

A worked orphan-risk register entry (illustrative, for a developer's periodic risk report; the format adapts the quarterly Risk Innovation Planner^18^):

Risk: emotional reliance on companion-mode assistants (orphan categories: social trends; loss of agency). *Value threatened*: customer wellbeing and autonomy (customers); youth mental health (community); litigation exposure and trust (enterprise). *Current status*: the subject of company research and now of litigation,^37^ yet tracked in no frontier safety framework; company attention is discretionary and unversioned. *Why not managed within the framework*: no agreed metric; harms accrue per user, below the severity floor. *Register action*: consolidate existing internal signals into one named, board-visible indicator. *Owner*: [named executive]. *This quarter*: commission an external review of the emerging clinical evidence; add reliance-pattern flags to usage dashboards; brief the board risk committee. *Review*: next cycle.

The entry does not commit the organization to quantify, or even to mitigate. It commits it to look, on the record — converting ad hoc attention into a disclosed and owned line item whose removal would itself be visible.

Box 2 | Three tests the argument must pass

The response proposed here comes from practice, and it should be tested like a hypothesis. Three falsifiable propositions would do so — two using data that become available on a known schedule. Incident-to-category coding rules should be fixed in advance and applied blind to the hypothesis.

Test 1 — De-listed risks bite. Risks removed from safety frameworks (persuasion, 2025; unconditional pause commitments, 2026) will be implicated in major AI-attributable incidents and enterprise costs at rates no lower than tracked risks, on a severity-weighted comparison, over 2026–2029. Two complications must be built into the coding rules. Tracked catastrophic risks are rare by design, and they are also actively mitigated — so low incident

rates in tracked categories could reflect successful treatment rather than sound selection. Orphaned-category incidents and costs should therefore also be judged against a materiality threshold fixed in advance; the hypothesis fails if incidents in tracked categories outpace those in orphaned ones.

Test 2 — The wedge narrows from the voluntary side. If risk selection is incentive-driven, enforcement supplies the incentive to internalize: categories a firm must answer for anyway should migrate into its safety framework within one to two revision cycles after EU enforcement begins (August 2026). Because a narrowing wedge is ambiguous between genuine re-selection and cosmetic convergence, the test also codes whether newly absorbed categories acquire thresholds and pre-committed responses, not merely mention. Persistence of the two-document structure past 2028 would count against the incentive-driven account of selection, though not against the case that risks are being orphaned.

Test 3 — Registers change scope, or their capture is visible. Developers adopting orphan-risk registers or aperture logs will show measurably different subsequent framework-revision behavior from matched non-adopters. Indistinguishable behavior would indicate ritual adoption — which would itself confirm the audit-society worry while counting against the register as a remedy.

Failure of the first test would count against the value-lens account itself; failure of the second, against the incentive-driven account of risk selection; the third bears on the remedy alone.

References

1. OpenAI. *Preparedness Framework (Beta)* (OpenAI, 18 December 2023); https://cdn.openai.com/openai-preparedness-framework-beta.pdf (accessed 3 July 2026).
2. OpenAI. *Preparedness Framework, Version 2* (OpenAI, 15 April 2025); https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf (accessed 3 July 2026).
3. OpenAI. *Frontier Governance Framework* (OpenAI, May 2026); https://cdn.openai.com/pdf/e37d949b-8c9f-4d76-b99e-4272f4631a7e/openai-frontier-governance-framework.pdf (accessed 3 July 2026).
4. California State Legislature. Senate Bill No. 53, Transparency in Frontier Artificial Intelligence Act, Stats. 2025, ch. 138 (2025); https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53 (accessed 3 July 2026).
5. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* L 2024/1689 (12 July 2024); https://eur-lex.europa.eu/eli/reg/2024/1689/oj (accessed 3 July 2026).
6. European Commission. *General-Purpose AI Code of Practice* (European Commission, 10 July 2025); https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (accessed 3 July 2026).
7. Anthropic. *Responsible Scaling Policy*, versions 1.0 (19 September 2023) to 3.3 (26 May 2026); changelog and archived versions at https://www.anthropic.com/rsp-updates (accessed 3 July 2026).
8. Anthropic. Sharing our compliance framework for California's Transparency in Frontier AI Act. *Anthropic* https://www.anthropic.com/news/compliance-framework-SB53 (19 December 2025) (accessed 3 July 2026).
9. Schaffelder, M. & Murray, M. Analyzing SSF requirements in the GPAI Code of Practice: a case study using Anthropic's Frontier Compliance Framework. *SaferAI* https://www.safer-ai.org/analyzing-ssf-requirements-in-the-gpai-code-of-practice-a-case-study-using-anthropics-frontier-compliance-framework (20 March 2026)

(accessed 3 July 2026).

10. Meta Platforms, Inc. *Annual Report on Form 10-K for the Fiscal Year Ended December 31, 2025* (US Securities and Exchange Commission, filed 29 January 2026); https://www.sec.gov/Archives/edgar/data/1326801/000162828026003942/meta-20251231.htm (accessed 3 July 2026).

11. Alphabet Inc. *Annual Report on Form 10-K for the Fiscal Year Ended December 31, 2025* (US Securities and Exchange Commission, filed 5 February 2026); https://www.sec.gov/Archives/edgar/data/1652044/000165204426000018/goog-20251231.htm (accessed 3 July 2026).

12. Google DeepMind. *Frontier Safety Framework*, versions 1.0 (May 2024) to 3.1 (April 2026); v3.0 announcement and version archive at https://deepmind.google/blog/strengthening-our-frontier-safety-framework/ (accessed 3 July 2026).

13. Meta. *Frontier AI Framework, Version 1.1* (Meta, 2025); https://web.archive.org/web/20250311152132/https://ai.meta.com/static-resource/meta-frontier-ai-framework/ (archive of 11 March 2025); and *Advanced AI Scaling Framework, Version 2* (Meta, 2026); https://ai.meta.com/static-resource/Meta_Advanced-AI-Scaling-Framework-v2 (accessed 3 July 2026).

14. UK Department for Science, Innovation and Technology. Frontier AI Safety Commitments, AI Seoul Summit 2024. *GOV.UK* https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024 (2024; updated 7 February 2025) (accessed 3 July 2026).

15. Slattery, P. et al. The AI Risk Repository: a comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. Preprint at https://arxiv.org/abs/2408.12622 (2024; database updated 2026) (accessed 3 July 2026).

16. Bengio, Y. et al. *International AI Safety Report 2026*. DSIT 2026/001 (UK Department for Science, Innovation and Technology, 2026); preprint at https://arxiv.org/abs/2602.21012 (accessed 3 July 2026).

17. Maynard, A. D., Oye, K. A., Scragg, M., Tripp, T. & Wolf, S. M. Successfully bridging innovation and application: exploring the utility of a risk innovation approach in the NSF Engineering Research Center for Advanced Biopreservation Technologies (ATP-Bio). *J. Law Med. Ethics* **52**, 553–569 (2024); https://doi.org/10.1017/jme.2024.126

18. Risk Innovation Nexus. Tools and resources for identifying and navigating orphan risks, including the *Risk Innovation Planner* and *Culminating Report* (Arizona State University, School for the Future of Innovation in Society & J. Orin Edson Entrepreneurship + Innovation Institute, 2017–2020); https://riskinnovation.org (accessed 3 July 2026).

19. Douglas, M. & Wildavsky, A. *Risk and Culture: An Essay on the Selection of Technological and Environmental Dangers* (Univ. California Press, 1982); https://www.jstor.org/stable/10.1525/j.ctt7zw3mr

20. Porter, T. M. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton Univ. Press, 1995); https://www.jstor.org/stable/j.ctt7sp8x

21. Karnofsky, H. Responsible Scaling Policy v3. *LessWrong* https://www.lesswrong.com/posts/HzKuzrKfaDJvQqmjh/responsible-scaling-policy-v3 (24 February 2026) (accessed 3 July 2026).

22. Koessler, L., Schuett, J. & Anderljung, M. Risk thresholds for frontier AI. Preprint at https://arxiv.org/abs/2406.14713 (2024) (accessed 3 July 2026).

23. Hackenburg, K. et al. The levers of political persuasion with conversational artificial intelligence. *Science* **390**, eaea3884 (2025); https://doi.org/10.1126/science.aea3884

24. Kasirzadeh, A. Two types of AI existential risk: decisive and accumulative. *Philos. Stud.* **182**, 1975–2003 (2025); https://doi.org/10.1007/s11098-025-02301-3

25. Power, M. *The Audit Society: Rituals of Verification* (Oxford Univ. Press, 1997).

26. Stelling, L. et al. Evaluating AI providers' frontier safety frameworks. Preprint at https://arxiv.org/abs/2512.01166 (2025) (accessed 3 July 2026).

27. Vaughan, D. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA* (Univ. Chicago Press, 1996).

28. Abbott, J. & Maynard, A. *AI and the Art of Being Human: A Practical Guide to Thriving with AI While Rediscovering Yourself in the Process* (Waymark Works Publishing, 2025).

29. International Organization for Standardization. *ISO 31000:2018 Risk Management — Guidelines* (ISO, 2018); https://www.iso.org/standard/65694.html; *ISO/IEC 23894:2023 Information Technology — Artificial Intelligence — Guidance on Risk Management* (ISO, 2023); https://www.iso.org/standard/77304.html (accessed 3 July 2026).

30. Kasperson, R. E. et al. The social amplification of risk: a conceptual framework. *Risk Anal.* **8**, 177–187 (1988); https://doi.org/10.1111/j.1539-6924.1988.tb01168.x

31. Stilgoe, J., Owen, R. & Macnaghten, P. Developing a framework for responsible innovation. *Res. Policy* **42**, 1568–1580 (2013); https://doi.org/10.1016/j.respol.2013.05.008

32. Maynard, A. D. & Garbee, E. Responsible innovation in a culture of entrepreneurship: a US perspective. In *International Handbook on Responsible Innovation: A Global Resource* (eds von Schomberg, R. & Hankins, J.) 488–502 (Edward Elgar, 2019).

33. Maynard, A. D. Why we need risk innovation. *Nat. Nanotechnol.* **10**, 730–731 (2015); https://doi.org/10.1038/nnano.2015.196

34. Heaven, W. D. Why Meta's latest large language model survived only three days online. *MIT Technology Review* (18 November 2022); https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/ (accessed 3 July 2026).

35. Mickle, T., Metz, C., Isaac, M. & Weise, K. Inside OpenAI's crisis over the future of artificial intelligence. *The New York Times* (9 December 2023); https://www.nytimes.com/2023/12/09/technology/openai-altman-inside-crisis.html (accessed 3 July 2026).

36. Leike, J. Posts on X (17 May 2024); https://x.com/janleike/status/1791498174659715494; archived at https://web.archive.org/web/20260609000957/https://x.com/janleike/status/1791498174659715494 (accessed 3 July 2026).

37. Raine v. OpenAI, Inc., No. CGC-25-628528 (Super. Ct. Cal., Cty. of San Francisco, filed 26 August 2025); complaint archived at https://web.archive.org/web/20260601012712/https://www.courthousenews.com/wp-content/uploads/2025/08/raine-vs-openai-et-al-complaint.pdf (accessed 3 July 2026).

38. OpenAI. *OpenAI Charter* (OpenAI, 2018); https://openai.com/charter (accessed 3 July 2026).

39. Maynard, A. D. The AI cognitive Trojan horse: how large language models may bypass human epistemic vigilance. Preprint at https://arxiv.org/abs/2601.07085 (2026) (accessed 3 July 2026).

40. Shaw, S. D. & Nave, G. Thinking — fast, slow, and artificial: how AI is reshaping human reasoning and the rise of cognitive surrender. Preprint at https://doi.org/10.31234/osf.io/yk25n_v1 (2026).