Harmful Information Management Practices in Frontier AI Development

Carson Ezell
Harvard University
cezell@college.harvard.edu

Ben Bucknall
University of Oxford
bucknall@robots.ox.ac.uk

Abstract

Information about risks from general-purpose AI (GPAI) systems is essential for effective risk management and oversight. Both developers and external stakeholders–such as auditors, regulators, insurers, and users–need credible risk evidence to make informed decisions. However, GPAI development firms ('firms') often control how such information is produced and shared, both internally and externally. Furthermore, they may have incentives to preclude, conceal, or distort information about AI risks. Drawing on historical patterns from industries like tobacco, pharmaceuticals, and chemicals, this paper categorizes four types of information management practices that can impede risk transparency and management: (1) *information generation*–influencing what research is conducted and how; (2) *information visibility*–controlling what information is shared internally and externally; (3) *perceived information credibility*–shaping how risk evidence is interpreted and trusted; and (4) *information acknowledgment*–avoiding or limiting recognition of risks. We also present seven categories of policy options to reduce the use or impact of these practices: improving scientific oversight, altering liability and immunity rules, externalizing risk assessments, limiting confidentiality protections, protecting parties reporting risk information (e.g., whistleblowers), expanding legal privileges, and mandating experimentation. For each, we highlight parallels from other sectors, potential benefits, and key policy design challenges. Together, these insights show how governance reforms could counter harmful information management and foster more reliable evidence for AI risk oversight.

1 Introduction

Information about risks from general-purpose AI (GPAI) systems–such as OpenAI's ChatGPT or Anthropic's Claude–is essential for effective risk management and oversight [Bommasani et al., 2025, Casper et al., 2025]. Employees of GPAI development firms ('firms') need this information to assess and mitigate risks from their systems. External parties–including auditors, insurers, users, and regulators–also need this information to assess risks and hold firms accountable.

However, firms have significant influence over how risk information is generated, shared, and framed–for example, choosing which evaluations to run on proprietary models, what findings to disclose, and how to characterize safety concerns. Strategic management of these information flows can impede risk mitigation [Kolt et al., 2024, METR, 2025b]. For example, avoiding certain tests to limit legal exposure can deprive internal teams and outside evaluators of the evidence they need to assess harmful capabilities.

These dynamics are not unique to AI. Across sectors–including tobacco, pharmaceuticals, and chemicals–companies have used similar tactics to downplay, conceal, or obscure product risks

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Regulatable ML.

[Wagner, 1996, White and Bero, 2010, Legg et al., 2021, Choo and Meyer, 2023, Schwarcz, 2023]. These practices have delayed efforts to protect public health and safety.

This paper overviews the types of information management practices in which GPAI development firms may engage to preclude, conceal, or distort evidence of AI risks. We draw on cases of other industries to illuminate these practices and, where applicable, identify public observations of GPAI development firms engaging in them. We then introduce seven categories of policies to reduce the use or negative impacts of these practices–for each, we include an illustrative example, provide an analogue to an implemented policy in another industry, and discuss key policy design challenges. We also describe more example policies from each category in Appendix B. We conclude by describing the need for more empirical work on firms' information management practices to inform policy analysis.

2 Practices

We use the term 'information management practices' to refer to firms' behaviors that increase the epistemic distance between stakeholders–including regulators, evaluators, insurers, the public, or firms' employees–and the actual risks posed by an AI system. This definition focuses on the informational consequences of actions rather than on the intention behind them. It allows us to objectively analyze behaviors that obscure risk, regardless of whether they stem from deliberate deception or alternative motivations. For example, practices implemented to improve security–such as siloing information within a GPAI development firm–can also prevent employees from identifying and addressing risks.

We organize information management practices into four categories based on where they intervene in the lifecycle of information: (1) *information generation*–influencing what research is conducted and how; (2) *information visibility*–controlling what information is shared internally and externally; (3) *perceived information credibility*–shaping how risk-related information is interpreted and trusted; and (4) *information acknowledgment*–avoiding or limiting recognition of known risks. These categories are also described with examples in Table 1. Within each category, we identify numerous tactics in which firms may engage. We also provide relevant examples of firms' behaviors for each tactic where they have been publicly reported. Notably, we are not making assumptions about firms' objectives or incentives in employing these tactics, nor attributing to them nefarious motivations. While we overview the main categories of tactics in this section, a more comprehensive list of tactics is included in Appendix A.

2.1 Information generation

Information generation tactics aim to influence what evidence about risks is created in the first place, ensuring that more favorable evidence is produced and unfavorable evidence is never created. In the pharmaceutical industry, companies have conducted clinical trials emphasizing a drug's benefits while designing them such that potential harms are statistically unlikely to be detected. A notorious case is Merck's painkiller Vioxx: Merck ran trials that were too short or otherwise underpowered to detect cardiac risks [Krumholz et al., 2007]. Internally, Merck scientists recognized the risks, but the company structured its publications to focus on other outcomes, delaying broader recognition of the danger [Krumholz et al., 2007]. Similarly, tobacco industry actors would fund research likely to produce evidence in support of product safety Bero et al. [1995], and lawyers would discourage or prevent researchers from pursuing work likely to create damaging evidence [Hanauer et al., 1995, Glantz et al., 1995].

2.1.1 Tactics

Influencing research directions and design: Developers may pursue research directions that are more likely to yield favorable evidence and less likely to yield damaging evidence. For example, they may deprioritize research on harmful capabilities elicitation. Developers may also manipulate research design, methods, and interpretation to avoid producing damaging evidence [Bero et al., 1995, Fletcher and Black, 2007, Wagner, 2022]. Google researchers reportedly must consult with legal and public relations teams before conducting research on topics that might reveal risks of the company's products [Schiffer, 2020a]–a policy which may have a deterrent effect on certain research directions.
Influencing model testing and evaluation: Developers may control the scope of tests, the methods

Category	Description	Examples from Other Industries	Example from GPAI Development Firms
Information Generation	Influencing what risk evidence is created to favor positive outcomes and avoid negative findings, including research direction, design, testing, and control of collaborators.	Merck's Vioxx trials designed too short/underpowered to detect cardiac risks; tobacco industry funding research likely to support product safety.	OpenAI reportedly stopped pre-release testing for risk of manipulating users; METR given limited time/access in safety assessments, constraining analysis.
Information Visibility	Controlling who can access or share risk evidence, limiting internal and external disclosures, resisting mandatory release, and publicizing only favorable findings.	DuPont hid internal findings for over 20 years showing PFOA caused health harms, never informing workers or the public.	Google DeepMind leadership reportedly blocked the release of a paper showing their models were less safe than a competitor's; xAI launched Grok without releasing a system card; OpenAI restricted internal information sharing about certain projects.
Perceived Information Credibility	Shaping how evidence is interpreted and trusted, including promoting favorable studies, discrediting critical ones, or reanalyzing data to counter negative results.	Tobacco industry's decades-long campaign to cast doubt on the smoking–disease link via e.g. industry-funded symposia, hiring expert consultants, and attacks on independent research.	OpenAI criticized the NYT for misuse, prompt manipulation, and cherry-picking when it provided evidence of models regurgitating articles.
Information Acknowledgment	Avoiding or limiting internal recognition of known risks, preventing documentation, ignoring external warnings, and abandoning risk-reduction research.	Tobacco firms routed external findings to lawyers to avoid broader awareness within industry and discouraged updates on certain research.	Microsoft engineer ordered to remove a LinkedIn post discussing harmful AI outputs; OpenAI disbanded teams focused on long-term safety research; OpenAI pressured leaving employees to sign non-disparagement agreements.

Table 1: Four categories of information management practices, with descriptions and examples from other industries and GPAI development. See main text for further description of examples and citations. See Appendix A for a complete list of tactics within each category.

used, and the interpretation of results–for example, when conducting pre-deployment safety evaluations. Prior to one release, OpenAI stopped testing its models for the risk of manipulating users without public explanation [Goldman and Kahn, 2025].

Influencing external collaborators or independent researchers: Developers can influence external researchers through funding decisions, constraints, and other incentives [Bero et al., 1995, Bero, 2005]. For instance, the evaluation group METR reports having had limited time and access during their safety assessments of OpenAI models, which constrained the depth of their analysis [METR, 2025a].

2.2 Information visibility

Information visibility tactics aim to control who can see the evidence that has been generated. In the chemical industry, DuPont was found to have hidden, for over 20 years, internal findings that perfluorooctanoic acid (PFOA) exposure posed serious health risks [Cone, 2005]. Company tests showed as early as 1981 that PFOA built up in human blood and caused birth defects–however, DuPont never informed its workers or external stakeholders of these findings [Andrews and Walker, 2015]. Instead, DuPont continued profiting from the chemical and avoided scrutiny until lawsuits and an EPA investigation forced disclosures.

2.2.1 Tactics

Limiting internal sharing of damaging information: Firms may restrict who internally can access or discuss certain risk-related information. For example, incident information might only be available to a small group of executives or lawyers [Bero, 2005]. Recent policy changes at OpenAI reportedly reduce internal information-sharing about some projects for security purposes, limiting what many staff can learn about certain safety issues [Riley, 2025].

Limiting external sharing of negative findings: Developers may curtail the publication of internal research or evaluation results that reflect poorly on their models, or subject such publications to lengthy and onerous review processes [Bero et al., 1995, Schwarcz, 2023]. Meta–a GPAI development firm–previously hid internal research about its social media platforms' negative effects on young users for many years [Wells et al., 2021]. Google DeepMind leadership also allegedly blocked a paper

from publication because it showed DeepMind's models were less safe than those of OpenAI [Morris and Heikkilä, 2025], and Google reportedly ordered an employee to retract a paper that painted its AI system in a critical light [Schiffer, 2020b]. Likewise, xAI departed from industry transparency norms by not releasing a system card for its *Grok* model upon launch [Zeff, 2025].

Resisting mandatory disclosure of evidence: If developers face litigation, they may attempt to keep damaging documents from becoming public by using protective orders, confidential settlements, or by asking courts to seal records–often citing trade secret protection [Wagner, 2022]. Developers facing lawsuits (including OpenAI and Meta) have attempted to seal extensive information about their models, with varying degrees of success [McKool Smith, 2025]. In one case, OpenAI only allowed opposing counsel–lawyers representing the New York Times–to view its training data in a monitored "secure room" without internet access, hampering plaintiffs' ability to analyze the evidence [Claburn, 2024]. The plaintiffs' lawyers further alleged that their data stored on a dedicated virtual machine (as part of their evidence) was erased by OpenAI engineers during discovery [Crosby and Lieberman, 2024].

Invoking legal privileges to hide risks: Companies may structure internal processes to keep certain risk information privileged and thus exempt from disclosure to regulators or courts. For instance, a developer might place lawyers in charge of incident investigations or claim that safety assessments were conducted under "anticipation of litigation," allowing them to invoke attorney–client privilege or work-product protection to shield those documents [Schwarcz, 2023]. Such tactics are common in the cybersecurity domain, where firms affected by data breaches try to protect evidence of vulnerabilities from being exposed [Kosseff, 2016, Woods et al., 2023, Schwarcz, 2023].

Publicizing favorable evidence: Conversely, developers often boost the visibility of evidence that portrays their systems in a positive light. They might highlight internal studies showing strong safety performance, issue press releases about benign or helpful model behaviors, or disproportionately use positive findings to bias regulators and courts [Bero, 2005].

2.3 Perceived information credibility

These tactics shape how information is interpreted and trusted by those who see it. A 1969 memo from the tobacco industry stated: "Doubt is our product since it is the best means of competing with the 'body of fact' that exists in the mind of the general public" [Huertas, 2022]. Tobacco companies organized their own research symposia and paid external experts to reinforce their scientific positions and attack independent research revealing health risks [Bero, 2005]. These tactics eroded public trust in legitimate science and delayed consensus on tobacco's dangers.

2.3.1 Tactics

Promoting favorable evidence misleadingly: Firms may share favorable findings through ostensibly trustworthy channels. For example, they might write company blog posts or white papers describing their model safety, or sponsor academic conferences and publications, to give more credibility to studies with favorable conclusions [Bero, 2005].

Undermining the credibility of critical studies: Conversely, developers can attempt to discredit researchers who publish damaging findings or to cast doubt on the findings themselves. They may make biased or frivolous methodological criticisms, or pressure authors to retract critical papers to discredit the findings [Shudtz et al., 2009]. For example, when the New York Times produced evidence of OpenAI models regurgitating their articles, OpenAI called out the New York Times for 'misuse' of their models, prompt manipulation, and cherry-picking [OpenAI, 2024].

Reinterpreting or countering negative evidence: Firms might seek access to the raw data behind an unfavorable study–through subpoenas or public records requests–and then reanalyze it in an effort to refute the original conclusions. They may also run new evaluations or fund new studies intended to contradict or cast doubt on the original, damaging findings [White and Bero, 2010, Wagner, 2022].

2.4 Information acknowledgment

These tactics involve limiting internal awareness or official recognition of evidence. For example, tobacco companies controlled the channels through which external researchers could share their findings with the company, using lawyers to receive such information rather than–for example–industry scientists or management [Hanauer et al., 1995]. They also made it clear to external scientists that they were not interested in updates on particular research to prevent receipt of such findings

[Hanauer et al., 1995]. Similarly, when companies are exposed to a cyber breach, they often ask incident investigators to avoid producing or sending them a final report summarizing their findings, or to leave out opinions and recommendations [Schwarcz, 2023]. Often, evidence related to breaches is restricted to a small control group within the company, preventing technical staff from accessing key information [Schwarcz, 2023].

2.4.1 Tactics

Avoiding documentation of risks: Firms may discourage producing written documentation about problems. For example, they may adopt policies or maintain a culture that steers employees away from formally documenting identified risks. In some cases, companies may even edit internal reports or external audit results to remove or soften discussions of risks.
Controlling employees' statements about risks: Firms may also impose guidelines on what employees or executives can publicly say about safety issues, and constrain what external collaborators (such as auditors) are allowed to publish. For example, when a Microsoft engineer made a LinkedIn post that revealed harmful outputs from a Microsoft AI product, he was promptly ordered to remove the post [Field, 2024a]. Furthermore, OpenAI formerly pressured departing employees to sign non-disparagement agreements by threatening vested equity [Piper, 2025].
Ignoring or siloing outside warnings: Some developers may choose not to collect externally generated information about risks, or to route such information through legal departments, keeping industry engineers and researchers unaware of problems. By siloing external evidence of harms, companies limit internal recognition of those issues.
Abandoning lines of research that signal risk: To avoid implicitly acknowledging a serious risk, a developer might quietly shut down or conceal internal research efforts that focus on studying it. For example, while its motivations are unclear, OpenAI has disbanded multiple teams focused on long-term AI safety and risk research [Field, 2024b].

3 Policy Options

Policy interventions can alter developers' incentives to engage in information management practices or mitigate their negative consequences. Drawing on lessons from other industries, this section analyzes seven categories of policy interventions. For each category, we describe the intended effects, an illustrative example, an analogous policy in another domain, and key policy design challenges. We also provide more examples of policies within each of the categories in Appendix B.

3.1 Improve science oversight

Policies can strengthen oversight of the scientific process around AI risk research and evaluation to ensure rigor and transparency. Such oversight can make it more difficult for industry actors to use biased methods, hide negative findings, or mislead the public about their results.

Example policy: Research registries. Regulators could establish an AI risk research registry–similar to clinical trial registries–where developers must pre-register safety-critical studies or evaluations and later disclose the results [Shudtz et al., 2009]. A public registry makes it harder to simply not publish problematic findings, since omissions would be visible [Dickersin and Rennie, 2003]. The U.S. Food and Drug Administration Amendments Act of 2007 mandated the pre-registration of many clinical trials and the posting of their results [Califf et al., 2025]. This created a public record of each trial's existence and predefined endpoints before results are known. While there have been some limitations in terms of comprehensiveness and quality of reporting [Zarin et al., 2007, Viergever et al., 2014], the registry has improved transparency [Califf et al., 2025].

Policy design challenges: Scientific oversight measures can be costly to administer and place an administrative burden on researchers, ultimately slowing down the research process [Pham and Oh, 2021]. Furthermore, they may still leave opportunities for researchers to selectively report or distort findings. For example, researchers may conduct preliminary testing before reporting the existence of a study at all, or–in the absence of mandates–only comply with scientific oversight measures when findings are favorable [Pham and Oh, 2021].

Some forms of scientific oversight may also create barriers that discourage researchers from beginning certain projects. For example, researchers may be more hesitant to conduct evaluations or exploratory

Category	Brief Description	Cross-Industry Analogue	Example Implementation	Main Benefits for AI Risk Governance	Key Design Challenges / Risks
Improve science oversight	Strengthen rigor and transparency of AI risk research/evaluation.	Clinical trial registries (ClinicalTrials.gov)	Require preregistration of safety-relevant studies on frontier models.	Counters biased study design and selective non-publication; increases scrutiny.	Administrative burden; selective compliance; potential for "rubber-stamping."
Alter liability & immunity rules	Adjust liability to incentivize thorough testing and disclosure.	Vaccine Injury Compensation Program (1986)	Safe harbors for firms meeting testing/reporting standards.	Encourages proactive evidence generation and sharing with regulators.	Setting adequate standards; minimal compliance; potential to reduce incentives for ongoing safety improvements.
Externalize risk assessments	Require third-party or regulator-led model evaluations.	NTSB accident investigations.	Government-conducted AI incident investigations with public recommendations.	Reduces developer control over scope/visibility of testing; increases credibility.	Ensuring assessor independence/competence; avoiding narrow scope; ensuring internal testing.
Limit confidentiality protections	Narrow scope for trade secret/NDA claims over safety info.	TSCA toxicity data disclosure rules	Ban confidentiality over AI-related public safety information even if proprietary.	Improves access to safety evidence; reduces legal suppression.	Risk to legitimate IP; upstream evidence avoidance; enforcement challenges.
Protect parties reporting risk info	Shield whistleblowers and external critics from retaliation.	FAA/NASA Aviation Safety Reporting System.	Anonymous incident reporting with immunity from certain penalties.	Increases flow of negative findings; deters internal suppression.	Preventing frivolous/malicious claims and opportunism; avoiding defensive adaptations; preserving internal reporting channels.
Expand legal privileges	Create protected channels for sharing risk info internally/with oversight bodies.	Patient Safety and Quality Improvement Act (2005).	Privilege for safety reports shared with certified external parties.	Encourages documentation and frank reporting; reduces litigation fears.	Potential misuse to hide evidence; reduced public transparency; need for conditional privileges.
Mandate experimentation	Mandate creation/retention of risk-relevant evidence.	DuPont PFOA settlement–funded epidemiological panel.	Court-ordered independent research to resolve uncertainties.	Prevents plausible deniability; ensures data for post hoc analysis.	Box-ticking compliance; privacy concerns and cost burdens.

Table 2: Description of policy categories to mitigate information management practices, with cross-industry analogues, example implementations, main benefits, and key design challenges.

research that may produce damaging findings if there would be corresponding requirements or expectations to disclose this work [Pham and Oh, 2021]. More generally, there is a risk that external transparency requirements decrease what becomes known internally.

When scientific oversight reforms rely upon independent reviews, there is also a risk that reviewers are too closely aligned with industry interests. They may simply rubber-stamp biased studies, which could worsen the credibility problem.

3.2 Alter liability and immunity rules

Modifications to the liability landscape can incentivize developers to test for and/or disclose safety problems rather than hide them. For example, penalty defaults–where developers face liability for harms by default, unless they can show that they took adequate safety measures–create incentives to demonstrate safety [Yew and Hadfield-Menell, 2022]. Safe-harbor provisions, meanwhile, reward developers for thorough testing and transparency by reducing their legal exposure accordingly.

Example policy: Liability safe harbors. Developers could be granted reduced liability if they perform rigorous testing (or allow independent testing) and address any risks that are identified [Wagner, 1996]. For example, a company could receive immunity (or a cap on damages) for certain unforeseen harms if it can demonstrate that it followed industry-standard evaluation practices and transparently reported the results. As one example, amid legal uncertainty facing vaccine manufacturers, the National Vaccine Injury Compensation Program (VICP) provides an alternative

to traditional tort suits for vaccine-related injuries [Ridgway, 1999]. Vaccine manufacturers fund a compensation pool via an excise tax, and injured patients apply for redress from this pool rather than suing manufacturers directly [Cook and Evans, 2011]. In exchange, manufacturers are largely shielded from liability, reducing their incentives to hide evidence of harms.

Policy design challenges: A key challenge is setting the right standards for immunity or presumptions. If the bar for a safe harbor is set too low (e.g., if minimal testing qualifies a company for immunity), developers might do only the bare minimum, leaving significant risks unexamined. Safe harbor programs would also need a mechanism to verify that companies actually performed high-quality tests (e.g. regulatory audits or documentation requirements). Otherwise, companies might be tempted to design superficial tests that are easy to pass. Another risk of liability safe harbors is that they reduce firms' incentives to innovate to improve product safety–indeed, this poor incentive exists for vaccine manufacturers [Lior, 2019]. Conversely, overly harsh penalty defaults or unclear testing mandates might chill innovation by making deployment legally perilous.

3.3 Externalize risk assessments

Frontier AI models can be assessed by independent parties (third-party experts or government agencies), with findings made available to relevant stakeholders or the public. The evidence that third parties produce is less likely to be skewed in the firm's favor than internal studies, and firms have a reduced ability to control the narrative. The presence of outside scrutiny might also deter firms from biasing or misrepresenting internal findings, since an external party might expose those gaps.

Example policy: Government testing and investigations. Agencies could be empowered to conduct their own evaluations of AI models (for instance, by having the authority to subpoena model access or training data). In the event of AI-related incidents or complaints, regulators might perform independent investigations to identify the root causes and contributing factors [Knake et al., 2021]. In aviation, the independent National Transportation Safety Board (NTSB) investigates major aircraft accidents, separate from both the industry and the regulator (the FAA). The NTSB publicly releases detailed analyses of crashes and issues safety recommendations [National Transportation Safety Board, 2025]. The NTSB provides credible, third-party information that leads to industry-wide safety improvements.

Policy design challenges: If independent evaluators are constrained by firms (e.g. given only limited access or narrow evaluation scopes), they will struggle to rigorously evaluate risks. This may result in incomplete findings that underestimate risks. Furthermore, ensuring the independence and quality of third-party assessors is crucial. Potential risks include conflicts of interest (e.g. close relationships with firms), 'forum shopping' where markets for external evaluations create perverse incentives [Hadfield and Clark, 2023], and uneven competence across external evaluators. Close engagement with developers is key to gain the tacit knowledge needed to run rigorous evaluations, but it can also partially compromise the evaluator's independence. There is also a risk of creating a false sense of security–if a model passes an independent audit, a company might be tempted to reduce its own internal testing, which might nevertheless surface unique issues.

3.4 Limit confidentiality protections

Developers' ability to use broad confidentiality claims to shield safety-relevant information–for example, via trade secret claims or sealed settlements–can be limited. These limitations would make it more difficult for developers to quietly continue risky practices.

Example policy: Narrow scope of trade secrecy claims. Regulators or courts could apply a higher bar for granting trade secret or confidential business information (CBI) protection to safety-relevant information. For instance, firms might be required to justify any claim that specific training data or model details need secrecy when those details bear on public safety risks [Shudtz et al., 2009]. Analogously, the Toxic Substances Control Act (TSCA) prevents companies from claiming confidentiality over health and safety studies pertaining to toxic substances [McGarity and Shapiro, 1980].

Policy design challenges: Overly aggressive limits on confidentiality could expose legitimate trade secrets and undermine firms' competitive advantages, potentially slowing innovation. Policymakers would need to target the rules carefully–e.g., differentiating between disclosing information about

model risks versus disclosing the model's proprietary architecture or code. Additionally, firms might respond by shifting what they consider "safety evidence" into formats that still qualify for protection–for instance, treating certain evaluations as privileged legal analyses. Strong oversight and enforcement would be needed to prevent abuse of any remaining confidentiality provisions. Limiting confidentiality might also discourage firms from analyzing problems or producing documentation, skipping written analyses or not investigating certain issues to avoid leaving a paper trail.

3.5 Protect parties reporting risk information

Protections and channels can be improved for individuals who present information about AI risks–whether they are industry insiders or independent researchers. When insiders or external critics are shielded from retaliation, intimidation, or suppressive lawsuits (i.e., SLAPPs), they may be more likely to surface and/or disclose damaging findings [Shudtz et al., 2009]. New state-level legislation in California has created some whistleblower protections for employees of AI companies [Tran, 2025].

Example policy: Whistleblower protections. Employees of GPAI development firms who report safety issues (whether internally or to regulators) should be legally protected from retaliation. The U.S. Federal Aviation Administration supports an anonymous reporting system operated by NASA, where pilots, air traffic controllers, and others can submit incident reports without fear of enforcement action. Under this program, if someone files a report about a safety lapse or error, the FAA waives penalties for the reporter (provided the incident was not intentional or criminal) [Aviation Safety Reporting System, 2021]. This immunity encourages individuals to report near-misses and hazards candidly, greatly expanding the information available to improve aviation safety. Policies could also include financial rewards or bounties if the reported information leads to enforcement action, similar to whistleblower reward programs in financial fraud.

Policy design challenges: These protections can trigger defensive adaptations. For instance, companies may alter contractual provisions or confidentiality terms to discourage sharing damaging information [Moberly, 2018], tighten internal silos, limit access to sensitive information, or avoid documenting potential problems altogether to reduce the risk of leaks. In extreme cases, firms might preemptively withdraw from certain testing or external collaborations to keep risk-related information from reaching protected reporters.

Broad immunity or blanket confidentiality carve-outs for whistleblowers may enable malicious or frivolous disclosures, or the release of legitimately sensitive information (e.g., proprietary technical details unrelated to safety). Poorly designed external whistleblower programs can also create opportunistic incentives if rewards or protections are easily accessed without strong good-faith requirements [Vega, 2012].

Reporting mechanisms and incentives that make external channels more attractive than reporting to firms (e.g., internal whistleblowing or coordinated disclosure) can also discourage the latter, depriving firms of early opportunities to address problems before they escalate [Vega, 2012]. In addition, regulators may be overwhelmed by low-priority or marginally relevant reports, straining investigative capacity and diverting attention from higher risk cases [Bloch-Wehba, 2024].

3.6 Expand legal privileges

Firms may hesitate to share some information with key stakeholders (internally or externally) if doing so would waive legal privileges which shield that information from discovery [Kosseff, 2016, Schwarcz, 2023]. Expanding certain legal privileges and immunities for communications about AI risks could encourage firms to document and share problems, by allowing them to share safety information with designated parties (e.g., regulators, insurers, auditors, or internal teams) without fear that those communications will later be used against them in court. For example, an engineer might be allowed to write a detailed incident report if that report is privileged, whereas they might have only given an oral debrief otherwise.

Example policy: Privilege incident-related communications with certain third parties. Legal privileges could be extended to cover information shared with outside auditors or insurance providers pertaining to incidents [Schwarcz, 2023]. If companies know that what they tell their insurer or auditor about model vulnerabilities will not later be exposed in court, they may be more honest

and forthcoming, leading to improved risk modelling. In healthcare, the Patient Safety and Quality Improvement Act (PSQIA) created a privilege for hospitals that share data on medical errors with certified Patient Safety Organizations. The information reported is shielded from use in lawsuits, which encourages hospitals to candidly report and analyze mistakes without fear of liability [Office for Civil Rights, 2025].

Policy design challenges: Overly broad protections can create loopholes for companies to hide evidence by routing it through a "privileged" process, while narrow and poorly targeted protections might not change behavior. Protections would need to have exceptions (e.g. evidence of criminal conduct) and/or conditions (e.g. implementing risk mitigations). Expanding privileges would also reduce public or plaintiff-side visibility of evidence, but the trade-off may be acceptable if it leads to companies, regulators, and certain third-parties knowing about the risks and addressing them. The net effect on information management practices depends on how privileges are structured and the mechanisms through which information flows incentivize and facilitate risk mitigations.

3.7 Mandate experimentation

Developers can be obligated to produce new scientific evidence about risks in some cases–for example, when there are significant uncertainties about risks that impede judicial or regulatory decision-making [Lahav, 2020]. Relatedly, developers may be required to preserve evidence and artifacts to enable later experimentation [Schwarcz et al., 2022]. Mandates to fill information gaps leave fewer possibilities for developers to hide evidence or sustain ambiguity. While emerging regulations, such as the EU AI Act, create some testing obligations for developers [European Commission, 2025], there are not robust processes for regulators to demand subsequent experimentation when developers' testing regimes leave ambiguity.

Example policy: Court-ordered research. If a lawsuit or investigation reveals significant uncertainty about a risk, a firm could be ordered to fund and facilitate independent research to resolve the uncertainty [Lahav, 2020, Wagner, 2022]. For example, if an AI system is alleged to have caused harm in a novel way, the court could mandate the developer to sponsor a study focused on that failure mode. In a 2005 settlement over contamination from a Teflon-related chemical (PFOA, or C8), DuPont agreed to fund an independent scientific panel to study the chemical's health effects on the affected community [Frisbee et al., 2009]. The panel conducted extensive epidemiological research, and in 2012 it concluded that exposure to the chemical was linked to several diseases. These findings were then used to guide medical monitoring and spurred broader regulation of perfluorinated chemicals (PFAS) [Lahav, 2020].

Policy design challenges: Mandating evidence generation can lead to a compliance mentality–companies might treat the requirement as a box-checking exercise, doing the bare minimum to satisfy regulators rather than genuinely probing for issues. Required tests need to be well-designed to produce useful information. Data retention also raises concerns around privacy (e.g. storing sensitive user data) and cost, especially for retaining large volumes of data. Policymakers would need to calibrate these requirements to avoid excessive burden while facilitating investigations where necessary.

4 Limitations and Future Work

In this work, we described how governance reforms could counter harmful information management practices and foster more reliable information for AI risk oversight. We emphasize that the information management practices we describe in this paper–and in many cases, attribute to GPAI development firms–may be used without the intention to preclude, hide, or distort information. For example, limits on internal information visibility may be imposed as information security measures. Furthermore, we note that if the policies we describe are poorly designed or implemented, they may be ineffective or produce negative consequences.

In future work, we plan to conduct semi-structured interviews with GPAI industry employees, auditors, regulators, and other stakeholders to better map out the extent to which information management practices are occurring, the key motivations driving them, and their consequences for risk management and oversight. This empirical foundation will be crucial for determining which (if any) policy reforms are most needed and for designing those policies to target harmful practices.

A Tactics

Table 3: List of tactics to conceal information about harms. We draw on Glantz et al. [1995], Hanauer et al. [1995], Bero et al. [1995], Bero [2005], Fletcher and Black [2007], White and Bero [2010], Legg et al. [2021], Wagner [2022], Schwarcz et al. [2022].

Type	Objective	Category	Tactic
Information generation	Increase quantity of favorable evidence	Influence scope of testing/evaluation and/or research directions	Produce evidence of safety to influence litigation outcomes or regulatory decision-making
			Favor tests/evaluations and/or research directions likely to produce desirable evidence
			Selectively fund research likely to produce favorable evidence of safety
		Influence study design and methods for research or testing/evaluation	Use biased methods to produce more favorable findings
		Influence testing and research interpretation and analysis	Selectively use data in analysis to support favorable findings
			Choosing statistical tests that support desired results
			Control how data is interpreted to produce favorable interpretations
		Influence external collaborators/researchers and their work	Fund external audits or assessments that are likely to produce favorable findings
			Create incentives for external researchers to produce favorable findings
	Reduce quantity of damaging evidence	Influence testing scope and research directions	Steer researchers away from work likely to produce damaging evidence
			Avoid tests/research likely to produce evidence of certain risks
			Require lawyer approval before pursuing research topics likely to produce damaging evidence
			Refuse or reduce funding for projects that take risks seriously
		Influence test and research design and methods	Involve lawyers in oversight of research design and methods
			Use biased methods to produce more favorable findings
		Influence test and research interpretation and analysis	Selectively use data in analysis to support favorable findings
			Choose statistical tests that support desired results
			Obscure or remove damaging data
			Control how data is interpreted to produce favorable interpretations
		Influence external collaborators/researchers	Avoid external audits or assessments that are likely to produce damaging findings

Continued on next page

Table 3 – *Continued from previous page*

Type	Objective	Category	Tactic
			Create incentives for external researchers to not produce damaging findings
			Use threats and intimidation to limit critical external research
			Limit scope or resources availalble for external firms' work (e.g., evaluators)
Information visibility	Reduce visibility of damaging evidence	Protect damaging evidence from discovery (attorney client privilege and/or work product doctrine)	Mediate communications with external firms through lawyers
			Give legal departments control over documents
			Restrict direct communications between company employees and external stakeholders (e.g., evaluators)
			Make legal departments responsible for contracting/coordinating with external parties (e.g., auditors)
		Limit external sharing of damaging evidence	Lawyers control external dissemination of research
			Refuse to share damaging documents with third parties
			Refuse to share raw data externally
			Disincentivize researchers from publicly sharing damaging findings
			Establish lengthy/onerous review processes before publishing/externally sharing papers
			Prevent the publication/dissemination of research with damaging evidence
			Invoke trade secrecy to avoid sharing damaging evidence
			Invoke other confidentiality rules (eg CPSA 6b) to avoid sharing damaging evidence
			Protective orders to reduce visibility of evidence
			Seal records in trials to avoid damaging evidence from being disclosed
			Use settlement agreements to prevent damaging evidence from being disclosed (e.g., NDAs)
			Employ nondisclosure clauses when funding or collaborating with external researchers
		Limit internal sharing of damaging evidence	Lawyers control internal dissemination of research
			Limit communication within firm about damaging evidence to control group (eg management and lawyers)
	Increase visibility of favorable evidence	Directly share favorable findings	Bypass publication/peer review processes
			Widely disseminate favorable information
			Discredit critics of favorable research

Continued on next page

Table 3 – *Continued from previous page*

Type	Objective	Category	Tactic
			Share findings directly with policymakers and regulators
		Alter information sharing channels	Increase reliance on industry information in policymaking
Perceived information credibility	Improve favorability of perceptions of evidence	Enhance perceived credibility of favorable research	Sponsor symposia or conferences to publish without peer review
			Conceal involvement of industry in research
			Selectively publish works with favorable findings
		Promote misleading positions	Make misleading statements about findings
			Reinforce misleading evidence in internal company documents/memos
	Reduce credibility of damaging evidence	Criticize damaging research	Offer accusations of scientific misconduct, conflicts of interest, and/or unfair biases
			Encourage researchers to retract papers with damaging evidence
			Directly criticize damaging research (e.g., press releases)
			Hire and encourage experts and/or consultants to criticize research (e.g., public statements, letters to the editor, etc.)
		Reassess external research	Use third-party subpoenas and public-record requests for data and research records
			Conduct or fund novel research to refute critical findings
			Misinterpret external research data
Information acknowledgment	Reduce recognition of negative evidence	Conceal efforts to design safer products	Avoid efforts to design safer products
			Keep efforts to design safer products secretive
		Prevent recognition of external damaging evidence	Avoid referencing research that contradicts the industry position (including through policies or legal advice)
		Control reception of external reports of damaging evidence	Make legal departments responsible for receiving external communications
			Limit report intake from external stakeholders (e.g., reports from auditors or bug bounty programs)
	Reduce documentation of damaging evidence and statements	Limit damaging statements made externally	Limit discussion of harms/risks by industry scientists/executives
			Establish guidelines for making public statements about risks
		Reduce documentation of damaging statements	Establish a culture of avoiding documentation/acknowledgment of harms
			Craft internal policies designed to ensure that damaging information is not recorded in the first place

Continued on next page

Table 3 – *Continued from previous page*

Type	Objective	Category	Tactic
			Avoid policies/culture that limit internal stakeholders from offering risk-related opinions/recommendations
			Discuss sensitive issues orally (not in writing)
		Control statements made by external parties	Prevent external parties (e.g., auditors) from issuing recommendations
			Avoid constructing external reports (e.g., evaluation reports) with damaging evidence
			Establish policies that limit former employees from making damaging statements
		Edit existing documents	Edit documents produced internally to remove discussions of damaging evidence or opinions/recommendations
			Destroy evidence of risks and/or evidence of their existence
			Have lawyers or regulatory affairs teams review and/or edit manuscripts prior to submission/publication or external release
			Have lawyers approve documents before internal circulation
			Edit reports from external audits, assessments, and evaluations to remove damaging evidence or opinions/recommendations

B Additional Example Policies

This appendix provides further illustrative examples of potential policies within each of the seven policy categories discussed in Section 3. These examples are intended to complement the main text by expanding on possible variations, adaptations from other domains, and novel mechanisms that could be considered in AI governance.

B.1 Improve Science Oversight

Facilitate replication: Developers could be encouraged or required to allow third parties to attempt to replicate their risk-related research findings. This might involve sharing model access, datasets, and methods with independent researchers to enable replication studies.

Independent peer review: Companies could submit internal risk research for review by an external scientific panel or board of experts. These independent reviewers would verify that the methods are sound and that results (including negative results) are fully reported, before the research is released or used for regulatory purposes.

Research oversight boards: Governments might charter independent review boards for high-stakes AI research [Wagner, 1996]. These boards could set standards for evaluation methods and verify that companies follow best practices. As an example of such practices, the Environmental Protection Agency enforces GLP standards for non-clinical studies submitted in regulatory applications (e.g., for pesticides and industrial chemicals) [US EPA, 2013]. These rules dictate standards for study protocols, data recording, and records retention, helping ensure a baseline level of quality and integrity for industry-submitted safety data.

B.2 Alter Liability and Immunity Rules

Penalty defaults for not testing. Lawmakers could establish a presumption of defect or causation if a developer fails to conduct reasonable safety tests on an AI system [Yew and Hadfield-Menell, 2022]. In practice, this would mean that if a harm occurs and the company did not perform standard tests, the court would assume the AI was defective and the company was negligent. Such a rule would make avoiding tests more legally risky than conducting them.

B.3 Externalize Risk Assessments

Third-party model audits or certifications: Regulators could require that frontier AI models undergo assessment by an independent, accredited auditor or evaluation service before deployment. The results of these audits (e.g., compliance certifications or risk ratings) could be shared with regulators, insurers, or even the public.

Industry-funded evaluation research: One approach is to establish a pool of funds (financed by AI developers, perhaps proportionally to their R&D spending) to support independent academic research on AI risks [White and Bero, 2010]. Grants from this pool could enable external scientists to evaluate company models or investigate emerging safety issues without facing industry pressure.

B.4 Limit Confidentiality Protections

Sunshine in litigation: Courts could refuse to seal court records or settlement agreements that contain evidence of significant public risks posed by AI systems. Similarly, judges could limit protective orders so that evidence of hazards uncovered in discovery cannot be completely kept from regulators or the public.

Expiry or carve-outs: Confidentiality protections for AI information could be time-limited, or explicit carve-outs could be introduced to ensure that data about safety risks must eventually be disclosed, even if some technical details (e.g. source code) remain protected.

NDA limitations: Lawmakers could declare certain nondisclosure agreement (NDA) clauses unenforceable–specifically, those that would prevent individuals from reporting safety issues to regulatory authorities or from warning users about documented risks.

B.5 Protect Parties Reporting Risk Information

Anonymous reporting channels: Regulators could set up secure portals for AI practitioners to anonymously submit concerns about unsafe practices or incidents. These channels would reduce fear of career damage by keeping the source confidential and shielding the whistleblower from reprisal.

Anti-SLAPP and researcher protections: Independent scientists, journalists, or others who expose AI risks could be shielded from intimidatory legal tactics. For example, stronger anti-SLAPP laws (to quickly dismiss "Strategic Lawsuits Against Public Participation") could deter companies from suing researchers or critics simply to silence them. Additionally, a new cause of action could allow researchers to sue if they are subjected to bad-faith legal demands (like baseless subpoenas for data) aimed at harassment or suppression [Shudtz et al., 2009].

B.6 Expand Legal Privileges

Safe sharing with regulators: Create a legal privilege for communications about AI risks between a company and regulators or oversight bodies. For example, if a company self-reports a problem to a regulator and shares a detailed analysis, that report might be shielded from discovery in civil litigation. This encourages openness with regulators without increasing lawsuit exposure [Schwarcz et al., 2022].

Internal investigative privilege: Extend legal privileges to cover certain aspects of internal safety investigations. For instance, if a designated internal safety team investigates an incident and produces a report, that report could be deemed protected (not admissible in court) as long as the company shares it with the relevant regulator and addresses the findings. This way, companies can investigate thoroughly without fearing that the report will be used adversarially by plaintiffs.

B.7 Require Evidence Generation

Regulatory information orders: Analogous to product recalls, regulators could issue orders requiring an AI developer to conduct specific tests or analyses when evidence emerges of a new kind of risk. The results would need to be reported back to the regulator by a deadline, and appropriate mitigations taken.

Data and log retention mandates: Regulations could require developers to retain certain data that might be critical for post hoc analysis of harms. For instance, companies could be required to securely store AI system interaction logs and records of training data or model versions for a set period. This ensures that if an incident occurs, investigators have access to the necessary evidence. Similarly, regulators might require that companies archive snapshots of model parameters (or at least maintain the ability to reproduce a given model version) for forensic analysis.

References

David Andrews and Bill Walker. Poisoned Legacy, May 2015. URL https://www.ewg.org/research/poisoned-legacy.

Aviation Safety Reporting System. Immunity Policies. Advisory Circular 00-46F, April 2021. URL https://asrs.arc.nasa.gov/overview/immunity.html?utm_source=chatgpt.com.

Lisa Bero, Deborah E. Barnes, Peter Hanauer, John Slade, and Stanton A. Glantz. Lawyer Control of the Tobacco Industry's External Research Program: The Brown and Williamson Documents. *JAMA*, 274(3):241–247, July 1995. ISSN 0098-7484. doi: 10.1001/jama.1995.03530030061035. URL https://doi.org/10.1001/jama.1995.03530030061035.

Lisa A. Bero. Tobacco industry manipulation of research. *Public Health Reports*, 120(2):200–208, 2005. ISSN 0033-3549. doi: 10.1177/003335490512000215. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1497700/.

Hannah Bloch-Wehba. The Promise and Perils of Tech Whistleblowing. *Northwestern University Law Review*, 118(6):1503–1562, April 2024. ISSN 0029-3571. URL https://scholarlycommons.law.northwestern.edu/nulr/vol118/iss6/2.

Rishi Bommasani, Sanjeev Arora, Jennifer Chayes, Yejin Choi, Mariano-Florentino Cuéllar, Li Fei-Fei, Daniel E. Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, Arvind Narayanan, Alondra Nelson, Emma Pierson, Joelle Pineau, Scott Singer, Gaël Varoquaux, Suresh Venkatasubramanian, Ion Stoica, Percy Liang, and Dawn Song. Advancing science- and evidence-based AI policy. *Science*, 389(6759):459–461, July 2025. doi: 10.1126/science.adu8449. URL https://www.science.org/doi/abs/10.1126/science.adu8449. Publisher: American Association for the Advancement of Science.

Robert M. Califf, Tracy L. Cutler, Hilary D. Marston, and Ann Meeker-O'Connell. The importance of ClinicalTrials.gov in informing trial design, conduct, and results. *Journal of Clinical and Translational Science*, 9(1):e42, January 2025. ISSN 2059-8661. doi: 10.1017/cts.2025.9. URL https://www.cambridge.org/core/journals/journal-of-clinical-and-translational-science/article/importance-of-clinicaltrialsgov-in-informing-trial-design-conduct-and-results/125B3C69C8923DC03550090EBB7E7A12.

Stephen Casper, David Krueger, and Dylan Hadfield-Menell. Pitfalls of Evidence-Based AI Policy, April 2025. URL http://arxiv.org/abs/2502.09618. arXiv:2502.09618 [cs].

Chun Wei Choo and Marco Meyer. Information misbehavior: How organizations use information to deceive. *Journal of the Association for Information Science and Technology*, 74(9):1081–1085, 2023. ISSN 2330-1643. doi: 10.1002/asi.24804. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24804. _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24804.

Thomas Claburn. OpenAI will show secret training data to copyright lawyers, September 2024. URL https://www.theregister.com/2024/09/26/openai_training_data_author_copyright_case/.

Marla Cone. DuPont Settles Charges That It Hid Toxic Risk Data, December 2005. URL https://www.latimes.com/archives/la-xpm-2005-dec-15-fi-dupont15-story.html. Section: Business.

Katherine M. Cook and Geoffrey Evans. The National Vaccine Injury Compensation Program. *Pediatrics*, 127(Supplement_1):S74–S77, May 2011. ISSN 0031-4005. doi: 10.1542/peds.2010-1722K. URL https://doi.org/10.1542/peds.2010-1722K.

Ian Crosby and Steven Lieberman. Re: The New York Times Company v. Microsoft Corporation, et al., 23-cv-11195; Daily News, L P, et al. v. Microsoft Corp., et al., 1:24-cv-3285, November 2024. URL https://storage.courtlistener.com/recap/gov.uscourts.nysd.612697/gov.uscourts.nysd.612697.328.0.pdf.

Kay Dickersin and Drummond Rennie. Registering Clinical Trials. *JAMA*, 290(4):516–523, July 2003. ISSN 0098-7484. doi: 10.1001/jama.290.4.516. URL https://doi.org/10.1001/jama.290.4.516.

European Commission. Third Draft of the General-Purpose AI Code of Practice, March 2025. URL https://digital-strategy.ec.europa.eu/en/library/third-draft-general-purpose-ai-code-practice-published-written-independent-experts.

Hayden Field. Microsoft engineer warns company's AI tool creates violent, sexual images, ignores copyrights, March 2024a. URL https://www.cnbc.com/2024/03/06/microsoft-ai-engineer-says-copilot-designer-creates-disturbing-images.html. Section: Technology.

Hayden Field. OpenAI disbands another safety team, as head advisor for 'AGI Readiness' resigns, October 2024b. URL https://www.cnbc.com/2024/10/24/openai-miles-brundage-agi-readiness.html. Section: Technology.

Robert H. Fletcher and Bert Black. Spin in Scientific Writing: Scientific Mischief and Legal Jeopardy Defining Scientific Misconduct. *Medicine and Law*, 26(3: Scientific Misconduct):511–526, 2007. URL https://heinonline.org/HOL/P?h=hein.journals/mlv26&i=535.

Stephanie J. Frisbee, A. Paul Brooks, Arthur Maher, Patsy Flensborg, Susan Arnold, Tony Fletcher, Kyle Steenland, Anoop Shankar, Sarah S. Knox, Cecil Pollard, Joel A. Halverson, Verónica M. Vieira, Chuanfang Jin, Kevin M. Leyden, and Alan M. Ducatman. The C8 Health Project: Design, Methods, and Participants. *Environmental Health Perspectives*, 117(12):1873–1882, 2009. ISSN 0091-6765. doi: 10.1289/ehp.0800379. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2799461/.

Stanton A. Glantz, Deborah E. Barnes, Lisa Bero, Peter Hanauer, and John Slade. Looking Through a Keyhole at the Tobacco Industry: The Brown and Williamson Documents. *JAMA*, 274(3):219–224, July 1995. ISSN 0098-7484. doi: 10.1001/jama.1995.03530030039032. URL https://doi.org/10.1001/jama.1995.03530030039032.

Sharon Goldman and Jeremy Kahn. OpenAI updated its safety framework—but no longer sees mass manipulation and disinformation as a critical risk, April 2025. URL https://www.yahoo.com/news/openai-updated-safety-framework-no-190931446.html.

Gillian K. Hadfield and Jack Clark. Regulatory Markets: The Future of AI Governance, April 2023. URL http://arxiv.org/abs/2304.04914. arXiv:2304.04914 [cs].

Peter Hanauer, John Slade, Deborah E. Barnes, Lisa Bero, and Stanton A. Glantz. Lawyer Control of Internal Scientific Research to Protect Against Products Liability Lawsuits: The Brown and Williamson Documents. *JAMA*, 274(3):234–240, July 1995. ISSN 0098-7484. doi: 10.1001/jama.1995.03530030054034. URL https://doi.org/10.1001/jama.1995.03530030054034.

Aaron Huertas. Doubt is their product: How Big Tobacco, Big Oil, and the Gun Lobby market ignorance, June 2022. URL https://www.fastcompany.com/90759059/doubt-is-their-product-how-big-tobacco-big-oil-and-the-gun-lobby-market-ignorance.

Rob Knake, Adam Shostack, and Tarah Wheeler. Learning from Cyber Incidents, November 2021. URL https://www.belfercenter.org/sites/default/files/pantheon_files/files/publication/Learning%20from%20Cyber%20Incidents.pdf.

Noam Kolt, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin Esvelt, Gillian K. Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B. Sandbrink, and Thomas Woodside. Responsible Reporting for Frontier AI Development, April 2024. URL http://arxiv.org/abs/2404.02675. arXiv:2404.02675 [cs].

Jeff Kosseff. The Cybersecurity Privilege, August 2016. URL https://papers.ssrn.com/abstract=3225782.

Harlan M Krumholz, Joseph S Ross, Amos H Presler, and David S Egilman. What have we learnt from Vioxx? *BMJ : British Medical Journal*, 334(7585):120–123, January 2007. ISSN 0959-8138. doi: 10.1136/bmj.39024.487720.68. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1779871/.

Alexandra D. Lahav. The Knowledge Remedy, June 2020. URL https://papers.ssrn.com/abstract=3627868.

Tess Legg, Jenny Hatchard, and Anna B. Gilmore. The Science for Profit Model—How and why corporations influence science and the use of science in policy and practice. *PLOS ONE*, 16(6): e0253272, June 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0253272. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0253272. Publisher: Public Library of Science.

Anat Lior. AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy, August 2019. URL https://papers.ssrn.com/abstract=3446115.

Thomas O. McGarity and Sidney A. Shapiro. The Trade Secret Status of Health and Safety Testing Information: Reforming Agency Disclosure Policies. *Harvard Law Review*, 93(5):837–888, 1980. ISSN 0017-811X. doi: 10.2307/1340420. URL https://www.jstor.org/stable/1340420. Publisher: The Harvard Law Review Association.

McKool Smith. AI Infringement Case Updates: January 20, 2025, January 2025. URL https://www.mckoolsmith.com/newsroom-ailitigation-6.

METR. Details about METR's preliminary evaluation of OpenAI's o3 and o4-mini, April 2025a. URL https://metr.github.io/autonomy-evals-guide/openai-o3-report/.

METR. What should companies share about risks from frontier AI models? *METR Blog*, June 2025b. URL https://metr.org/blog/2025-06-27-risk-transparency/.

Richard Moberly. Confidentiality and Whistleblowing, 2018. URL https://papers.ssrn.com/abstract=3345644.

Stephen Morris and Melissa Heikkilä. DeepMind slows down research releases to keep competitive edge in AI race. *Financial Times*, April 2025. URL https://www.ft.com/content/2ee1ffde-008e-4ea4-861b-24f15b25cf54.

National Transportation Safety Board. Safety Recommendations, 2025. URL https://www.ntsb.gov/investigations/Pages/safety-recommendations.aspx?utm_source=chatgpt.com.

Office for Civil Rights. Understanding Confidentiality of Patient Safety Work Product, 2025. URL https://www.hhs.gov/hipaa/for-professionals/patient-safety/index.html. Last Modified: 2025-07-01T14:12:15-0400.

OpenAI. OpenAI and journalism, February 2024. URL https://openai.com/index/openai-and-journalism/.

Michel Tuan Pham and Travis Tae Oh. Preregistration Is Neither Sufficient nor Necessary for Good Science. *Journal of Consumer Psychology*, 31(1):163–176, 2021. ISSN 1532-7663. doi: 10.1002/jcpy.1209. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcpy.1209. _eprint: https://myscp.onlinelibrary.wiley.com/doi/pdf/10.1002/jcpy.1209.

Kelsey Piper. OpenAI NDAs: Leaked documents reveal aggressive tactics toward former employees | Vox. *Vox*, March 2025. URL https://web.archive.org/web/20250314110303/https://www.vox.com/future-perfect/351132/openai-vested-equity-nda-sam-altman-documents-employees.

Derry Ridgway. No-Fault Vaccine Insurance: Lessons from the National Vaccine Injury Compensation Program. *Journal of Health Politics, Policy and Law*, 24(1):59–90, February 1999. ISSN 0361-6878. doi: 10.1215/03616878-24-1-59. URL https://doi.org/10.1215/03616878-24-1-59.

Duncan Riley. OpenAI tightens internal security amid fears of IP theft by Chinese AI rivals, July 2025. URL https://siliconangle.com/2025/07/08/openai-tightens-internal-security-amid-fears-ip-theft-chinese-ai-rivals/. Section: AI.

Zoë Schiffer. Google reportedly asked employees to 'strike a positive tone' in research paper, December 2020a. URL https://www.theverge.com/2020/12/23/22197760/google-sensitive-topics-review-research-papers-timnit-gebru.

Zoë Schiffer. Google fires prominent AI ethicist Timnit Gebru, December 2020b. URL https://www.theverge.com/2020/12/3/22150355/google-fires-timnit-gebru-facial-recognition-ai-ethicist.

Daniel Schwarcz. How Privilege Undermines Cybersecurity. *Articles*, January 2023. URL https://scholarship.law.umn.edu/faculty_articles/1040.

Daniel Schwarcz, Josephine Wolff, and Daniel W. Woods. How Privilege Undermines Cybersecurity. *Harvard Journal of Law & Technology (Harvard JOLT)*, 36:421, 2022. URL https://heinonline.org/HOL/Page?handle=hein.journals/hjlt36&id=428&div=&collection=.

Matt Shudtz, Rena Steinzor, and Wendy Wagner. Saving Science from Politics: Nine Essential Reforms of the Legal System, May 2009. URL https://progressivereform.joanpiedra.com/publications/savingscience805/.

Lam Tran. Governing Frontier AI: California's SB 53, October 2025. URL https://www.lawfaremedia.org/article/governing-frontier-ai--california-s-sb-53. Publisher: Lawfare.

OECA US EPA. Good Laboratory Practices Standards Compliance Monitoring Program, July 2013. URL https://www.epa.gov/compliance/good-laboratory-practices-standards-compliance-monitoring-program.

Matt Vega. Beyond Incentives: Making Corporate Whistleblowing Moral in the New Era of Dodd-Frank Act Bounty Hunting. *Connecticut Law Review*, January 2012. URL https://digitalcommons.lib.uconn.edu/law_review/181.

Roderik F. Viergever, Ghassan Karam, Andreas Reis, and Davina Ghersi. The Quality of Registration of Clinical Trials: Still a Problem. *PLOS ONE*, 9(1):e84727, January 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0084727. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0084727. Publisher: Public Library of Science.

Wendy Wagner. When a Corporation's Deliberate Ignorance Causes Harm: Charting a New Role for Tort Law. *DePaul Law Review*, 72:413, 2022. URL https://heinonline.org/HOL/Page?handle=hein.journals/deplr72&id=425&div=&collection=.

Wendy E. Wagner. Choosing ignorance in the manufacture of toxic products. *Cornell L. Rev.*, 82:773, 1996. URL https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/clqv82&section=35. Publisher: HeinOnline.

Georgia Wells, Deepa Seetharaman, and Jeff Horwitz. Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show. *WSJ*, September 2021. URL https://www.wsj.com/tech/personal-tech/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739. Section: Tech.

Jenny White and Lisa A. Bero. Corporate Manipulation of Research: Strategies are Similar Across Five Industries. *Stanford Law and Policy Review*, 21:105, 2010. URL https://heinonline.org/HOL/Page?handle=hein.journals/stanlp21&id=107&div=&collection=.

Daniel W. Woods, Rainer Böhme, Josephine Wolff, and Daniel Schwarcz. Lessons lost: incident response in the age of cyber insurance and breach attorneys. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, pages 2259–2273, USA, August 2023. USENIX Association. ISBN 978-1-939133-37-3.

Rui-Jie Yew and Dylan Hadfield-Menell. A Penalty Default Approach to Preemptive Harm Disclosure and Mitigation for AI Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 823–830, Oxford United Kingdom, July 2022. ACM. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534130. URL https://dl.acm.org/doi/10.1145/3514094.3534130.

Deborah A. Zarin, Nicholas C. Ide, Tony Tse, William R. Harlan, Joyce C. West, and Donald A. B. Lindberg. Issues in the Registration of Clinical Trials. *JAMA*, 297(19):2112–2120, May 2007. ISSN 0098-7484. doi: 10.1001/jama.297.19.2112. URL https://doi.org/10.1001/jama.297.19.2112.

Maxwell Zeff. OpenAI and Anthropic researchers decry 'reckless' safety culture at Elon Musk's xAI, July 2025. URL https://techcrunch.com/2025/07/16/openai-and-anthropic-researchers-decry-reckless-safety-culture-at-elon-musks-xai/.