

The Safety Exposure Paradox: How Frontier AI Safety Research May Amplify Cyber Risk in Financial Infrastructure

Abena Primo, Azubike Okpalaeze, Al Amin, and Rohan Thompson

Department of Computer Science
Huston-Tillotson University
Austin, Texas, USA

Department of Business
Huston-Tillotson University
Austin, Texas, USA

acprimo@htu.edu

adokpalaeze@htu.edu

aamin@htu.edu

rrthompson@htu.edu

<https://doi.org/10.68414/BDBH8187>

ABSTRACT

Frontier artificial intelligence laboratories now publish detailed accounts of how their models behave under adversarial pressure: red-team transcripts, capability thresholds, interpretability findings, and responsible-scaling frameworks. These disclosures are widely treated as safety infrastructure, built to keep dangerous capabilities contained and visible to regulators. This paper argues that the same disclosures can function as a second, less examined kind of infrastructure: a map of where a model's offensive ceiling sits, available to anyone willing to read it. Using a comparative case study of the published capability-disclosure frameworks of Anthropic, OpenAI, and Google DeepMind, three labs operating under a shared 2024 Seoul Summit safety mandate, this paper develops a concept called the Safety Exposure Paradox, in which the transparency that makes frontier AI governable also narrows the search space for adversaries probing model behavior. The paper then traces this dynamic into financial infrastructure, where fraud detection, anti-money-laundering screening, identity verification, and algorithmic trading increasingly depend on AI systems whose upstream governance, training data, and safety ceilings sit outside any single institution's control. Drawing on financial-stability research, dual-use disclosure

literature, and recent reporting on AI-enabled synthetic identity fraud, the paper proposes a tiered disclosure model with concrete implementation steps, and a set of policy recommendations for regulators and fintech institutions managing inherited AI risk. The aim is not to argue against safety disclosure, but to show that disclosure decisions carry a financial-sector externality that current governance frameworks do not yet price in.

Keywords: frontier AI, safety disclosure, Safety Exposure Paradox, fintech cybersecurity, responsible disclosure, dual-use AI, financial infrastructure risk, AI governance

INTRODUCTION

Banks no longer think of artificial intelligence as a backend convenience. It sits underneath fraud screening, anti-money-laundering (AML) monitoring, credit decisioning, algorithmic trading, and the identity checks that decide whether a new account opens or gets flagged. None of that is new to people who work in fintech. What is newer, and less discussed in the fintech literature, is that the AI systems doing this work are themselves the subject of an entirely separate governance conversation happening upstream, inside a small number of frontier AI laboratories, largely out of financial institutions' view.

Those laboratories have, to their credit, built increasingly elaborate safety apparatus around their models: red-teaming programs, interpretability research, capability evaluations, and public scaling policies that describe, in real detail, what a model can and cannot yet do safely. Anthropic is the clearest example of this approach. Its Responsible Scaling Policy (RSP) defines AI Safety Levels, sets capability thresholds for cyber, biological, and autonomous-replication risk, and commits the company to publishing Risk Reports and Frontier Safety Roadmaps that describe, with redactions, how close its models are to crossing those thresholds (Anthropic, 2024, 2026).

This is, by most measures, a responsible way to build powerful technology. It is also, this paper argues, a way of publishing a map. A capability threshold is a statement about where a model's abilities sit relative to a danger line, and knowing where that line is drawn tells an attacker something useful even if the underlying weights stay private: how far current models are from a given offensive capability, and which capabilities are not yet considered a near-term concern. None of this requires breaching anyone's security; it only requires reading what frontier labs already publish.

This paper calls that dynamic the Safety Exposure Paradox: the same transparency mechanisms designed to make frontier AI governable can also compress the work an adversary has to do to understand a model's offensive ceiling. Importantly, this paper does not claim that capability disclosures have demonstrably caused specific financial-sector exploits; the causal link remains a plausible mechanism rather than a proven one. The contribution is to name and structure that mechanism, show it is visible across multiple labs' published governance documents, and connect it to fintech risk categories that are independently documented as growing. The paradox is not an argument against disclosure; withholding capability information from regulators and the public has its own, larger costs. The point is narrower: disclosure decisions made inside AI labs carry a financial-sector externality that current governance conversations rarely name, changing the threat environment for every downstream institution that depends on AI for fraud defense, identity verification, or trading infrastructure, whether or not those institutions had any say in the disclosure.

The remainder of the paper reviews relevant scholarship, explains the comparative case-study method applied to Anthropic, OpenAI, and Google DeepMind, works through that comparison, builds a four-part theoretical framework for the paradox, and closes with a tiered disclosure model with implementation steps, and policy recommendations for regulators, fintech institutions, and frontier AI developers.

LITERATURE REVIEW

Four bodies of scholarship bear on this paper's argument, and they do not currently talk to each other much: AI alignment and safety-evaluation research, cybersecurity disclosure theory, financial-stability research on AI dependency, and the dual-use literature that has grown up around frontier model capability testing.

AI Safety Transparency and Capability Evaluation

The dominant view in AI safety research treats transparency as close to an unqualified good. Red-teaming, interpretability work, and public capability evaluations are understood as mechanisms by which outside parties can hold a lab accountable for what its models can do. Anthropic's RSP is the most fully worked-out version of this approach among frontier labs, defining AI Safety Levels and Capability Thresholds for CBRN, cyber, and autonomy risk, and committing to publish Risk Reports describing how close current models sit to those thresholds

(Anthropic, 2026). The United States' National Institute of Standards and Technology has built a parallel, sector-neutral version of the same logic into its AI Risk Management Framework and 2024 Generative AI Profile, treating pre-deployment testing and incident disclosure as core risk-management functions (National Institute of Standards and Technology, 2023, 2024).

What this literature has not done is ask what a capability disclosure looks like from an adversary's position rather than a regulator's. A regulator reading a Risk Report wants reassurance the lab understands its own models; an adversary reading the same document wants a current estimate of how far a model's abilities sit from a danger threshold, useful intelligence about where to direct further probing. The alignment literature has not built this second reading into its account of what disclosure costs.

Cybersecurity Disclosure Theory and Its Limits for AI

Cybersecurity has a long-settled answer to a structurally similar problem: responsible disclosure. A researcher who finds a vulnerability privately notifies the vendor, who patches it before the finding becomes public, since a patched vulnerability loses most of its offensive value. Frontier-AI researchers building on this analogy have proposed adapting coordinated disclosure to dual-use model capabilities: O'Brien et al. (2024) describe a Coordinated Disclosure of Dual-Use Capabilities (CDDC) process routing findings through a government coordinator before public release, modeling cybersecurity's coordinated vulnerability disclosure norms.

The disanalogy is the more important point, however. A capability is not a bug, and a frontier model is not a single deployed system that gets patched once. A capability is a property of a class of trained systems; once an evaluation shows a technique elicits dangerous behavior in one model family, that knowledge generalizes to other families with similar architectures, in a way a software exploit does not generalize across unrelated codebases. Brenneis (2025) frames this as a dual-use-research-of-concern (DURC) problem borrowed from biosecurity, arguing AI research currently lacks the institutional layer that research ethics committees provide there. The cybersecurity disclosure model assumes a patchability that does not cleanly transfer to frontier AI, and that gap is exactly where the Safety Exposure Paradox lives.

Financial-Sector Dependency on AI Infrastructure

Financial-stability researchers have been more attentive than AI-safety researchers to the downstream risk of upstream concentration, though their focus has mostly been operational

rather than adversarial. The Financial Stability Board's 2024 report identifies third-party dependency and service-provider concentration as a primary systemic vulnerability, noting that financial institutions increasingly rely on a small number of specialized hardware, cloud, and pre-trained-model providers, which concentrates operational risk in ways difficult for any single regulator to see across institutions (Financial Stability Board, 2024). A 2025 follow-up report extends this with a case study on generative-AI supply chains, observing that vertical integration among AI and cloud providers is narrowing the pool of providers financial institutions can choose from (Financial Stability Board, 2025). The same 2024 report separately flags generative-AI-enabled fraud as a distinct vulnerability category, alongside, rather than connected to, its third-party dependency analysis.

This is where the present paper tries to do something the financial-stability literature has not yet done: connect the dependency question to the disclosure question. The FSB's framework treats AI vendor concentration as a risk because a vendor outage or failure could disrupt many institutions at once. It does not ask whether the vendor's own public safety disclosures might be quietly reshaping the threat landscape that all of those dependent institutions share, independent of any outage.

Adversarial Machine Learning and the Limits of Static Defense

A smaller but fast-growing technical literature documents how generative AI is already being used to defeat the specific controls fintech institutions rely on, particularly through synthetic identity construction and deepfake biometric spoofing that exploit the same generative capabilities frontier labs evaluate during pre-deployment testing (Feedzai, 2025; Datos Insights, 2025; PYMNTS, 2026). What this literature does not do is trace these capabilities back to the upstream question of where the generative models producing this fraud acquired the techniques they use, or whether published capability research played any role in that acquisition. That gap, between documenting that AI-enabled fraud is growing and explaining where the underlying capability came from, is the one this paper tries to close.

METHODOLOGY

This paper uses a qualitative, comparative case-study design, treating Anthropic, OpenAI, and Google DeepMind as what Yin (2018) would call a purposive sample of critical cases, chosen not for statistical representativeness but because they are the three frontier labs with the

most extensively documented, longest-running public capability-disclosure frameworks. All three were original signatories of the Frontier AI Safety Commitments at the 2024 AI Seoul Summit, where sixteen organizations agreed to publish a safety framework defining intolerable-risk thresholds and report publicly on breaches (UK Government, 2024). That shared starting point matters methodologically: the three frameworks were produced under a common institutional mandate, making differences between them a function of each lab's choices rather than different external requirements.

The three cases were selected for a second reason beyond the Seoul commitments: each has iterated its framework publicly enough times to show a visible trajectory rather than a single static snapshot. Anthropic has published five successive Responsible Scaling Policy versions between September 2023 and April 2026 (Anthropic, 2026). OpenAI has published two major versions of its Preparedness Framework, the second in April 2025, alongside model-specific system cards that apply the framework to each new release (OpenAI, 2025). Google DeepMind has published three versions of its Frontier Safety Framework between May 2024 and September 2025, most recently adding a Critical Capability Level for harmful manipulation (Google DeepMind, 2025b). Comparing frameworks that have each evolved over multiple iterations lets the analysis ask not only what each lab discloses, but how each lab has changed what it discloses, which is the more direct test of whether disclosure compresses over time in the way the paradox predicts.

The analysis proceeds in four stages. The first stage reviews the publicly available versions of all three frameworks, treating these as primary documents rather than secondary description. The second stage builds a structured comparison across four dimensions: the risk domains each framework tracks, the threshold structure each uses to define danger, each lab's stated default disclosure posture, and each lab's documented history of removing, adding, or redefining categories. The third stage places this comparison in conversation with the cybersecurity responsible-disclosure literature, asking where the disclosure logics of all three labs diverge from a patchable-vulnerability model. The fourth stage applies the resulting framework to five fintech functions, asking in each case what a capability disclosure in the Anthropic, OpenAI, or DeepMind style would and would not reveal to an adversary targeting that function.

This design narrows, but does not eliminate, the limitation of a single-case design. Three labs are still a small, non-random sample within an industry that includes other significant developers, including Meta and several labs outside the Seoul Summit’s signatory list, and the comparison cannot establish how common the Safety Exposure Paradox is across that wider population. The paper also has no access to non-public information about whether any specific capability disclosure has been used by a real attacker to compromise a real financial system; that evidence would require a confirmed incident or cooperation from threat-intelligence teams the present project did not have. What the comparative case study can do is show that the paradox is logically coherent, visible across multiple labs’ published governance documents, and that it maps onto fintech risk categories independently documented as growing.

RESULTS

This section reports what the comparative analysis found at each of the four stages described above.

Three Frameworks, One Underlying Logic

Table 1 summarizes the structural comparison across the three frameworks. All three track broadly overlapping risk domains, CBRN, cyber, and some form of autonomous or self-improvement risk, though only Anthropic and Google DeepMind currently maintain a standing persuasion or manipulation category; OpenAI moved persuasion outside its Preparedness Framework entirely in its 2025 revision, choosing to handle it through other internal mechanisms instead (OpenAI, 2025). All three use a tiered threshold structure rather than a single danger line, though the tiers differ: Anthropic’s AI Safety Levels (ASL-2 through ASL-4), OpenAI’s capability bands, and Google DeepMind’s Critical Capability Levels (CCLs) are not directly interchangeable, but each performs the same basic function of marking the point at which a lab commits to additional safeguards (Anthropic, 2024; OpenAI, 2025; Google DeepMind, 2025a, 2025b).

Table 1*Comparison of Frontier Capability-Disclosure Frameworks*

Dimension	Anthropic RSP	OpenAI Preparedness Framework	Google DeepMind FSF
Risk domains tracked	CBRN, cyber, autonomy/AI R&D, persuasion	Biological/chemical, cyber, AI self-improvement (persuasion handled outside framework)	CBRN, cyber, ML R&D, deceptive alignment, harmful manipulation
Threshold structure	AI Safety Levels (ASL-2–3)	Capability bands (originally Low/Medium/High/Critical ; Medium and Low later dropped)	Critical Capability Levels (CCLs)
Default disclosure posture	Minimize redactions in public Risk Reports	Publish results to “facilitate public awareness” of frontier capability	Publish framework and per-model evaluation outcomes
Documented trajectory	Five versions (2023–2026); categories disaggregated and added over time	Two major versions; low-risk bands removed as “not operationally involved”	Three versions (2024–2025); new CCL for manipulation added in v3.0

Anthropic’s RSP institutionalizes capability evaluation well beyond a one-time safety review, defining AI Safety Levels modeled loosely on the biosafety-level system for handling dangerous biological materials, with Capability Thresholds that trigger Required Safeguards once a model crosses them (Anthropic, 2024). Beginning with version 3.0, Anthropic committed to publishing Risk Reports covering all publicly deployed models and releasing a Frontier Safety Roadmap across security, alignment, safeguards, and policy (Anthropic, 2026), stating directly that it aims to minimize redactions in these public reports, accepting redaction only for legal, intellectual-property, public-safety, or privacy reasons, a default posture of maximal rather than minimal disclosure.

OpenAI’s Preparedness Framework follows a similar evaluate-then-disclose logic but structures its thresholds differently, and its revision history shows the disclosure choice working in the opposite direction from Anthropic’s. The original Beta framework defined four risk bands,

Low, Medium, High, and Critical; the April 2025 version 2 removed the bottom two, stating that Low and Medium “were not operationally involved” in any actual safety decision (OpenAI, 2025; Libertify, 2026). The same history shows OpenAI deploying its o1 model after evaluation found it reached Medium capability for biological and chemical severe harm and for persuasion severe harm simultaneously, with no additional safeguard required at that band (Libertify, 2026). Whatever the safety merits, it is also a disclosed data point: it tells any reader how far a real, deployed model’s capability sits from OpenAI’s High threshold in two categories at once.

Google DeepMind’s Frontier Safety Framework shows a third pattern: steady category expansion. It launched in May 2024 with four Critical Capability Levels covering autonomy, biosecurity, cybersecurity, and machine-learning R&D (Google DeepMind, 2024). Version 2.0 added Security Level recommendations and a deceptive-alignment domain for models that might resist shutdown (Google DeepMind, 2025a); version 3.0 added a wholly new CCL for harmful manipulation, building on internal research into manipulative mechanisms in generative models (Google DeepMind, 2025b). Each addition is a disclosed claim: it tells outside readers the lab now considers a specific capability close enough to dangerous to warrant a named threshold, where it previously did not.

What the Comparison Reveals About Disclosure as Signal

Read together, the three frameworks show the same mechanism expressing itself in three directions: Anthropic’s toward more granular categories, OpenAI’s toward fewer categories but with a documented case of deploying at a disclosed mid-tier capability band, and Google DeepMind’s toward new categories added as research reveals new harm mechanisms. Each still tells a sophisticated reader where a real, deployed model’s capabilities sit relative to a named danger line. This is where the comparison sharpens the paradox beyond a single-lab case: three labs, sharing a Seoul Summit mandate but arriving at different architectures, converge on the same property, naming a line and reporting distance to it, suggesting the exposure is structural to capability-threshold governance rather than one lab’s choice (UK Government, 2024).

None of this requires any of the three labs to publish anything resembling exploit code. The exposure described here is narrower: a threshold framework tells a sophisticated reader which capability categories are being tracked, and a multi-version changelog tells that reader how those categories are evolving, itself a roadmap for locating current systems relative to a danger line, without any document crossing into instructional detail.

Mapping the Comparison Onto Fintech Functions

The fourth stage applied this reading to five fintech functions. Fraud and AML detection depend on anomaly detection trained to recognize suspicious behavior; research already shows generative models producing synthetic transactions and identities realistic enough to survive KYC pipelines, with industry data indicating over half of measured 2025 fraud involved AI (Feedzai, 2025; PYMNTS, 2026). A concrete illustration: AI-generated synthetic identities built by combining real identity fragments with generative-model-produced photographs and voice samples have already been used to bypass major U.S. bank onboarding processes, with losses projected to reach \$58.3 billion by 2030 (PYMNTS, 2026). Digital identity verification faces a more severe version of the same problem, with deepfake biometric spoofing the fastest-growing fraud vector institutions report for the coming two years (Datos Insights, 2025). Algorithmic trading is exposed differently: reliance on similar AI models increases correlated trading behavior, compounding rather than offsetting any capability gap an adversary might exploit during a stress event (Financial Stability Board, 2024). Smart-contract auditing uses AI-assisted static analysis to find vulnerabilities; the same capability that helps an auditor find a flaw also helps an attacker find it first if they reach it sooner. None of these functions maps onto only one lab's framework, a reason fintech institutions cannot manage this exposure by tracking a single vendor alone.

DISCUSSION

The Safety Exposure Paradox: A Four-Part Framework

The case-study findings above can be organized into four interacting mechanisms, which together constitute what this paper calls the Safety Exposure Paradox.

Capability signaling is the most direct mechanism: a published threshold is a public statement about how close current models sit to an offensive ceiling, made so outside readers can verify it without needing model access themselves.

Adversarial mirroring follows from signaling: an attacker needs no privileged access to a lab's red-team transcripts to benefit from its evaluation work, only the same publicly trained model or a comparably capable alternative, and the published threshold tells the attacker roughly where to direct probing effort. The dual-use literature has named this dynamic for

biosecurity-style DURC concerns (Brenneis, 2025); this paper traces the same mechanism into a financial-sector context that literature does not address.

Governance escalation describes a secondary effect visible across all three labs studied: published capability findings do not stay inside the AI-safety community, but feed directly into legislative action. The Seoul Summit mandate that produced all three frameworks explains why: governments built the Frontier AI Safety Commitments specifically to extract public threshold disclosures that could inform regulation, and Anthropic cites the resulting influence on California's SB 53 and the EU AI Act's codes of practice as evidence the approach worked (UK Government, 2024; Anthropic, 2026). A disclosure made for accountability purposes can change compliance obligations or deployment availability for downstream institutions on a timeline they do not control, a dynamic structural to the disclosure regime rather than specific to any one lab's choices.

Inherited infrastructure risk is the mechanism connecting the first three to the fintech sector specifically. Financial institutions sit downstream of frontier AI governance decisions in the same structural position the Financial Stability Board has already flagged as a systemic vulnerability for unrelated operational reasons (Financial Stability Board, 2024, 2025): heavy reliance on a small number of AI and cloud providers. The same concentrated dependency that creates outage risk also creates disclosure risk, since a single upstream disclosure decision can change the threat environment for every dependent institution, with no mechanism for those institutions to weigh in.

Together, these four mechanisms describe a recursive problem rather than a simple trade-off: better safety evaluation produces more granular disclosure, more granular disclosure produces a clearer map of the offensive frontier, a clearer map narrows adversarial search costs, and narrower search costs eventually surface as new fraud, identity, or trading exploits that fintech institutions absorb without having created or understood the upstream decision behind them. It bears emphasizing that this chain of mechanisms is currently theoretical; confirming each link empirically, particularly the step from narrowed search costs to realized exploits, is an important direction for future research.

A Tiered Disclosure Model

If the paradox is real, the appropriate response is not to abandon capability disclosure, which would sacrifice the accountability benefits that motivated the alignment-transparency

literature in the first place. A more proportionate response is to tier disclosure by audience, an approach with precedent in the coordinated dual-use disclosure model O'Brien et al. (2024) propose for AI more broadly, adapted here specifically for fintech-relevant capability categories.

Tier 1 disclosures would remain fully public: general safety commitments, high-level governance structure, and aggregate risk posture, the level of detail that already appears in Anthropic's public-facing RSP summaries and serves the broad accountability function that motivates disclosure in the first place. Tier 2 disclosures would be restricted to regulated financial institutions and their primary supervisors, covering capability-threshold detail specific to fraud, identity-spoofing, and market-manipulation risk, distributed through an institutional channel comparable to existing financial-sector information-sharing arrangements rather than posted as a public PDF. Tier 3 disclosures would be limited to government coordinators and named security partners, covering the granular, near-term capability findings that the CDDC proposal already recommends routing through a dedicated coordinator before any public release (O'Brien et al., 2024), treating a sufficiently dangerous cyber-capability finding with roughly the handling discipline reserved for zero-day vulnerability intelligence in conventional cybersecurity practice.

Implementation of this model would proceed in three phases. In the near term (0–12 months), frontier AI labs and financial regulators should establish a joint working group to define which capability categories cross into Tier 2 territory, drawing on existing FS-ISAC and financial intelligence-sharing infrastructure to build a distribution channel. Financial institutions should simultaneously begin upstream AI dependency audits that catalog which frontier models underpin which critical functions. In the medium term (12–36 months), the working group should develop a standardized Capability Disclosure Template that labs submit to Tier 2 and Tier 3 recipients before publishing the public-facing version; government coordinators would review this template against their own threat-intelligence holdings and flag findings that meet Tier 3 handling thresholds. In the longer term, this template and handling protocol should be incorporated into the existing regulatory frameworks that govern model risk management at financial institutions, such as the U.S. Office of the Comptroller of the Currency's SR 11-7 model risk guidance and the EU AI Act's high-risk AI provisions, so that frontier-lab disclosure obligations are treated as a systemic risk input rather than a voluntary publishing choice.

This model does not eliminate the paradox; some signaling is unavoidable if disclosure is to serve any accountability function at all. It does, however, decouple the audience that needs granular threshold detail to do its job, regulators, sector-specific risk teams, and security partners, from the general public audience for whom that same detail offers little protective value and meaningful offensive value.

Policy Recommendations

1. Financial institutions should conduct upstream AI dependency audits that explicitly include exposure to capability-disclosure risk, not only vendor-outage risk, when assessing third-party AI providers.
2. Financial regulators should require AI provenance reporting from regulated institutions, documenting which frontier models support which critical functions, so that a capability disclosure affecting one model family can be traced to its downstream financial-sector footprint.
3. Regulators and institutions should encourage AI and cloud-provider diversification for safety-critical financial functions, reducing the systemic exposure concentrated reliance already creates under the Financial Stability Board's framework (Financial Stability Board, 2024, 2025), while frontier AI developers should adopt the risk-tiered disclosure structure proposed above, distinguishing public accountability disclosures from sector-specific threshold detail.
4. Frontier AI firms and financial regulators should establish a direct coordination channel, modeled on the CDDC proposal (O'Brien et al., 2024), for capability findings with foreseeable financial-sector application, handling cyber-relevant findings with disclosure discipline comparable to zero-day vulnerability intelligence rather than folding them into general-purpose safety reporting.

IMPLICATIONS, LIMITATIONS, AND FUTURE RESEARCH

This paper carries practical implications for three audiences. For fintech institutions, the most immediate implication is that vendor risk assessment should expand beyond operational resilience (uptime, data security, contractual SLAs) to include disclosure risk: what does this vendor publish about its models' offensive capabilities, and how might that publication change

the threat environment my fraud and AML controls face? For financial regulators, the implication is that model risk management frameworks, currently built around the risks a deployed model poses internally, should be extended to cover the systemic risk that concentrated reliance on a small number of frontier AI providers creates when any one of those providers makes a capability disclosure. For frontier AI developers, the implication is that disclosure decisions are not cost-free even when they serve legitimate accountability goals; the cost is borne by downstream financial institutions that had no seat at the disclosure table.

Sample size is this paper's central remaining limitation. Comparing three labs improves on a single case, showing the paradox surfacing across different disclosure philosophies rather than one company's idiosyncratic choices, but three labs remain a small, purposively chosen sample within a wider industry that includes Meta, Amazon, and several labs outside the original Seoul Summit signatory list. The paper also has no confirmed incident data linking a specific disclosure to a specific financial-sector exploit; the fintech risk categories discussed here are independently documented as growing, but the causal link to upstream disclosure remains a plausible mechanism rather than a demonstrated one. This is a genuine limitation that the authors acknowledge directly: the theoretical framework developed here is a necessary first step, but it does not substitute for empirical validation.

Future work could extend the comparison to a larger sample of signatories, or work with threat-intelligence teams to examine whether realized fraud incidents trace to the timing of major capability disclosures. A particularly valuable contribution would be a structured comparison between financial institutions that have diversified their AI infrastructure across multiple vendors versus those relying on a single frontier provider, testing whether disclosure concentration risk is reflected in measurable differences in fraud-event timing or severity. Researchers with access to confidential supervisory data at financial regulators would be well positioned to begin this work.

CONCLUSION

Frontier AI safety research is not purely protective, even though it is built and discussed almost entirely in protective terms. The same disclosure practices that let regulators and the public hold AI labs accountable also narrow what an adversary needs to learn before probing a model's offensive ceiling, and that narrowing does not stay confined to the AI industry. It reaches

financial institutions that had no role in the disclosure decision but now run fraud detection, AML screening, identity verification, and trading infrastructure on AI systems shaped by it.

For fintech specifically, this means inherited risk has a dimension that current operational-resilience frameworks do not yet capture. The Financial Stability Board has done useful work naming AI vendor concentration as a systemic vulnerability (Financial Stability Board, 2024, 2025), but that work treats the upstream provider mainly as a single point of operational failure. This paper's argument is that the same provider is also a single point of disclosure exposure, and that the two risks compound rather than sit side by side.

None of this argues for less transparency from frontier AI developers. It argues for transparency structured with an audience in mind: a disclosure aimed at reassuring the public and one aimed at protecting financial infrastructure are not the same document, and do not currently need to be. The tiered disclosure model and phased implementation steps proposed here offer a practical path toward that distinction. Building it into AI governance, rather than treating disclosure as an all-or-nothing choice, is the next step this framework points toward.

REFERENCES

- Anthropic. (2024, October 15). Anthropic's Responsible Scaling Policy.
<https://www.anthropic.com/news/anthropics-responsible-scaling-policy>
- Anthropic. (2026, February 24). Responsible Scaling Policy version 3.0.
<https://www.anthropic.com/news/responsible-scaling-policy-v3>
- Brenneis, A. (2025). Assessing dual use risks in AI research: Necessity, challenges and mitigation strategies. *Research Ethics*. <https://doi.org/10.1177/17470161241267782>
- Datos Insights. (2025). AI-powered synthetic identity attacks pose a novel risk to financial institutions.
<https://datos-insights.com/reports/ai-powered-synthetic-identity-attacks-pose-a-novel-risk-to-financial-institutions/>
- Feedzai. (2025, May 6). More than 50% of fraud driven by AI and hyper-realistic impersonations, but banks also use generative AI to fight back.
<https://www.feedzai.com/pressrelease/ai-fraud-trends-2025/>
- Financial Stability Board. (2024, November 14). The financial stability implications of artificial intelligence. <https://www.fsb.org/uploads/P14112024.pdf>
- Financial Stability Board. (2025, October 10). Monitoring adoption of artificial intelligence and related vulnerabilities in the financial sector.
<https://www.fsb.org/2025/10/monitoring-adoption-of-artificial-intelligence-and-related-vulnerabilities-in-the-financial-sector/>
- Google DeepMind. (2024, May 17). Introducing the Frontier Safety Framework.
<https://deepmind.google/blog/introducing-the-frontier-safety-framework/>
- Google DeepMind. (2025a, February 4). Updating the Frontier Safety Framework.
<https://deepmind.google/blog/updating-the-frontier-safety-framework/>
- Google DeepMind. (2025b, September 22). Strengthening our Frontier Safety Framework.
<https://deepmind.google/blog/strengthening-our-frontier-safety-framework/>
- Libertify. (2026, April 8). OpenAI Preparedness Framework 2025: What it actually guarantees about AI safety.
<https://www.libertify.com/interactive-library/openai-preparedness-framework-2025-safety-analysis/>

- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). <https://doi.org/10.6028/NIST.AI.100-1>
- National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). <https://doi.org/10.6028/NIST.AI.600-1>
- O'Brien, J., Smith, A., Adler, B., Anderljung, M., Bansal, A., Belfield, H., Bommasani, R., Christiano, P., Cooney, B., Garfinkel, B., Gleave, A., Heim, L., Hooker, S., Jordan, P., Kahn, L., Lewis, O., Mateos-Garcia, J., Murray, R., Sastry, G., Seger, E., Steinhardt, J., Thiel, D., Toner, H., & Wasil, A. (2024). Coordinated disclosure of dual-use capabilities: An early warning system for advanced AI. arXiv. <https://arxiv.org/abs/2407.01420>
- OpenAI. (2025, April 15). Preparedness Framework version 2. <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>
- PYMNTS. (2026, March 27). Synthetic identity fraud could reach \$58.3B: Are deepfakes driving a major industry blind spot? <https://www.pymnts.com/newswire-announcements/synthetic-identity-fraud-could-reach-58-3b-are-deepfakes-driving-a-major-industry-blind-spot/>
- UK Government. (2024, May 21). Frontier AI Safety Commitments, AI Seoul Summit 2024. <https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024>
- Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). SAGE Publications.