Incident Reporting Beyond Frontier Models

Monitoring and Preventing AI Harms in the Era of Open Models and AI Agents

October 11 , 2026

Dr. J. Nathan Matias
Dr. Baobao Zhang

Executive Summary

In 2026, even as New York finalized the Responsible AI Safety and Education (RAISE) Act, new developments in the efficiency of open weight models, the sophistication of AI agents, and the power of reinforcement learning expanded the scope and scale of AI risk beyond the law's reach. In this memo, we summarize these developments and offer suggestions for efforts to understand and limit critical AI incidents:

- Developments in AI agents and open models have expanded the actors who could be responsible for critical incidents, creating a gap in regulation and reporting requirements.
- Regulation based on compute thresholds cannot account for AI agents. More **effective risk criteria would account for autonomy, deployment context, and tool use**.
- Because open weight models hinder foundation model developers from observing incidents, **more effective incident reporting would span the deployment supply chain**.
- **Reinforcement learning models are incentivized to mislead testers and incident investigators**; science lacks reliable methodologies to minimize this problem.
- Because no standards exist for AI incidents, **transparency efforts will be unreliable without standards development** in **incident reporting** and **post-incident analysis** reporting.
- Since the RAISE act incentivizes some actors to avoid incident reporting, data collection could be improved with **infrastructure that lowers the cost of reporting** and strong **standards for documenting and minimizing conflicts of interest**.

Dr. J. Nathan Matias is a computer scientist and behavioral scientist and an Associate Professor at Cornell University who studies digital governance in groups and networks shaped by AI. The founder of the Citizens and Technology Lab, Nathan is co-author of the book *Auditing AI*, has advised many national governments and international organizations on AI policy, and has held positions at MIT, Harvard, Stanford, Columbia, and Princeton.

Dr. Baobao Zhang is a political scientist and applied statistician who researches and teaches AI governance as an Associate Professor at Syracuse University. She is the co-editor of the *Oxford Handbook of AI Governance* and has advised the U.S. government and the European Union on AI policy. She has held positions at Cornell, MIT, and Harvard.

Dr. Zhang writes in her individual capacity as a researcher at Syracuse University; she does not represent Syracuse University.

The **Citizens and Technology Lab** is a research lab at Cornell University that works alongside the public through community science, toward a world where digital power is guided by evidence and accountable to the public.

Contents

Executive Summary ii

1 Introduction 1

2 Regulating Based on Risk Tiers When Compute Thresholds Are Insufficient 3

3 Regulating AI Agents Across the Supply Chain of Deployment 5

4 Providing Reliable Claims About Model Function and Behavior In the World 6

5 Incident Reporting Standards for AI Agents 8

6 Preventing Incident Reporting Theater 10

7 Acknowledgments 12

References 13

1

Introduction

In the summer of 2026 when security researchers at OpenAI conducted a security evaluation with insufficient safeguards to prevent their models from compromising the systems of another AI company, the AI policy community became alarmed about the risks of AI models escaping their sandboxes. Foundation model developers quickly identified similar incidents at other firms that had not been previously discovered or disclosed. Yet an even more concerning event occurred in June that year when researchers at the University of Toronto demonstrated that publicly-available local AI models could power a worm that "adapts its strategy as it spreads from one device to the next" [6].

By early 2026, hundreds of thousands of AI agent tools had been deployed and downloaded, with half of downloaded agents designed to operate in unconstrained environments and most of those having action capacities such as browsing the web or controlling a computer. Many of these agents were designed to support finance and business operations [28]. In September, Meta released Muse, a system for agents that quickly reached over a million users [29]. These developments came alongside a wider, growing use of AI models in high risk areas such as healthcare [17, 37] and online child safety [4].

In the state of New York, the RAISE Act,[1] passed in 2026, governs critical harms from frontier AI models. The law defines frontier models on the basis of the monetary and computational resources required to train the model, formally designates institutions that develop these models, and requires developers to implement safety and security protocols, testing procedures, regular audits, reporting processes, and provides guidance on the conditions under which a frontier model should not be deployed.

The RAISE Act includes many common-sense elements that are essential to evidence-based regulation of any high-risk technology. When implemented well, transparency and incident reporting systems can alert key actors about critical risks in time for a coordinated response in time to contain and prevent disasters. Furthermore, if regulators intend to incentivize developers to innovate on safety technologies, then markets need access to reliable evidence on the safety of current systems, including information about critical failures [23]. Since such work requires human and technical capacity, we are encouraged by the creation of the industry-independent Office of Digital Innovation, Governance, Integrity and Trust (DIGIT).

Notably, the RAISE Act requires the reporting of safety incidents with frontier models, which are defined as an incident that “occurs in such a way that it provides demonstrable evidence of an

[1] https://legislation.nysenate.gov/pdf/bills/2025/A6453A

increased risk of critical harm." These incidents must be disclosed to the New York department of homeland security and emergency services within 72 hours of the developer learning of these incidents. The RAISE Act reporting requirements assume that frontier model developers will be able to learn about these incidents in a timely manner and that they have access to information about how the incidents occurred.

Since the RAISE Act and California's SB 53 were signed into law, AI model development and deployment has continued to grow both in capacity and efficiency. In November 2025, a developer published OpenClaw, a system that enables users to deploy AI agents in a wide variety of contexts. Beyond OpenClaw, agents have been given the capacity to direct human behavior, make purchases, and control decision-making systems, widening the range of possible harms arising from AI models.

In a further development, developers significantly improved the efficiency of AI models, permitting local models to run reliably on computers that cost hundreds to thousands of dollars. These local models, by design, do not report telemetry back to the foundation model creators. By June, cybersecurity researchers had demonstrated that these local agents could be used to create computer worms that spread through networks independently of their operators [6]. In September of 2026, NVIDIA acquired Hugging Face, a platform for distributing and testing local models, for $12.9 billion, accelerating an already-widespread practice of local model use. That same month, ConvAI innovations released Laya, a decision-making model that dispenses with text generation in order to operate much more efficiently than large language models, and which could be run on a mobile phone [20]. In September, Meta also released an AI Agent named Muse, which reached over three million weekly users by early October [29].

In this memo, we summarize general gaps in the RAISE act that derive from these step changes in efficiency and capacity. We also offer suggestions for ways that future laws and regulatory practices in New York State could account for these step changes.

2

Regulating Based on Risk Tiers When Compute Thresholds Are Insufficient

The RAISE Act currently uses training-compute as a proxy for AI model's risk: the law covers models trained with more than 10^{26} integer or floating-point operations. While we recognize that these criteria serve as a quantifiable and verifiable proxy for risk, empirical evidence has shown that training compute has become less correlated with risk as agentic models proliferate. As a result, we argue that the state should consider a **complementary assessment of models' risk that accounts for AI agents' autonomy, deployment context, and tool use**.

As model training becomes more efficient and models are deployed as agents, training compute threshold may no longer proxy for risks. For example, DeepSeek V4 Pro 0813, a highly capable Chinese open-weight model that was engineered for agentic workflows, required an estimated training compute of 9.7^{24} FLOP [2]. Developers could use computational methods like fine-tuning, model reuse, model expansion, and extra inference-time to create highly capable models without crossing training-compute thresholds [25].

Because the final version of the RAISE Act removed language covering models developed through knowledge distillation from frontier models, many highly capable open weight models are not subject to the law. We acknowledge that it may difficult to prove whether a model was developed using knowledge distillation and that developers of such models are unlikely to acknowledge this methodology since it violates the frontier AI companies' terms of service. This is another reason why compute thresholds are an insufficient means to measure risk.

More fundamentally, capability thresholds only protect society from risks that require a model to meet a threshold of competency. They cannot protect society from the risks of a sophisticated but incompetent AI model. For example, the development of a novel bioweapon requires a certain level of competence. But the misconfiguration of air traffic control does not.

For the reasons above, we recommend developing a complementary approach to classifying the risks associated with model use by considering these three dimensions of risk:

1. **Autonomy**: how much is the agent allowed to act without human oversight? Kasirzadeh & Gabriel (2026) grade autonomy from A.0 (no autonomy, the human directs every action) through A.3 (the human decides critical calls) to A.5 (full autonomy with no human oversight or control) [10].
2. **Deployment context**: where the agent is deployed and for what use cases? First we should consider what type of environment is the agent deployed in. Is the agent de-

ployed in a simulated environment that is constrained and resettable, such as an isolated testing environment? Is the agent operating in an environment where it can act only through human intermediaries? Or does it act directly in the physical environment without constraints, such as the open Internet [10]? Second, we should consider how risky is the use case, for instance whether the AI agent is used in law enforcement, critical infrastructure or education. For use case risk, one could consider the EU AI Act's four levels of risk (unacceptable, high, limited, and minimal) to classify the agent's potential to harm people [3].

3. **Tool use**: what tools do the agent have access to? These tools can be used to perceive the environment, reason (i.e., plan, analyze, or manage resources), or take action that directly affect their environment [21]. Concrete examples of "action" tools include logging into websites, using digital wallets, running computer code, interacting with humans over the phone or video chat, or controlling robotic arms. As part of assessing the type of tools the agent has access to, we should also consider the type of downstream permissions these tools have been granted. For example, we could ask if agents have "read-only" permission or are they allowed to "write" and influence their environment. According to a UK AI Security Institute analysis of 177,436 AI agent tools published on Model Context Protocol (MCP) server repositories between November 2024 and February 2026, the share of action tools in monthly downloads rose from 24% to 65%. Furthermore, half of the tools downloaded at the end of the study period allowed agents to operate in unconstrained environments like the open Internet [28].

We recognize that risks from AI agents are multi-dimensional and operationalizing such risks in legislation may require classifying agents into tiers. For instance, the highest risk tier would involve fully autonomous agents deployed in high-risk use cases that can use many tools with unconstrained "write" access. A concrete example for this highest risk tier would be fully autonomous agents used to manage the IT system of a hospital with the ability to execute code and edit the electronic medical records. Lower risk tiers may involve agents with lower degrees of autonomy deployed in minimal risk use cases that can use only a few tools with "read-only" access. A concrete example for such a lower risk tier would be agents used to spellcheck documents in a specified folder. Stanford University's AI Security Framework uses such a tier system (with three risk levels) to classify AI systems [27].

For future legislation, we recommend that AI agents in the highest risk tier face greater regulatory scrutiny, such as pre-deployment evaluation, incident reporting, and third-party audits. Because the information to assess risk requires comes from deployers (e.g., what use cases? what tools?), our recommendation acknowledges that regulation of AI agents needs to take place across the supply chain of deployment, as we explore in the next section.

3

Regulating AI Agents Across the Supply Chain of Deployment

Imagine a scenario where an AI agent is designed to coordinate information flows between a medical records vendor and a hospital system. The hospital requests that the vendor provide assurances about data privacy in compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA). That vendor could negotiate a contract with a foundation model developer to access compute from that developer to comply with their data privacy needs. Or they could deploy an open weight AI model on their own infrastructure that is completely separate from the systems of the firm that developed the foundation model.

Imagine then that this model exceeds its guardrails in a way that renders an entire hospital system unable to access reliable patient information. Such an occurrence would meet the standard of a critical incident under the RAISE Act. But under the RAISE act, the incident might not need to be reported, even if it involved a foundation model. In this section, we suggest that regulators **require incident reporting across the supply chain of deployment**.

The RAISE Act addresses a world where powerful models that are exclusively deployable in high-resource computing environments are those most likely to generate critical incidents. Under those circumstances, foundation model developers can provide model access as a service through APIs. This model-as-a-service infrastructure enables developers to maintain direct access to information about the behavior of their models as deployed by third parties.

These assumptions no longer apply. If the healthcare technology vendor deploys an open weight / open source model, the foundation model developer loses the ability to learn about critical incidents involving their model. As a result, they have nothing to report.

Since the RAISE act creates reporting obligations for model developers but not the vendors who deploy them, foundation model developers who produce open weight and open source models are disincentivized from discovering or reporting incidents related to local models, since that information exposes them to greater liability.

Even when deployers use models that send telemetry back to model developers, the RAISE act incompletely accounts for the structure of safeguards. Since foundation models are designed to conduct a wide range of tasks, model developers cannot anticipate all of the guardrails needed in deployment, many of which are deeply context-dependent [16, 18]. In the case of a healthcare incident, a detailed understanding of the incident would require investigations at all levels of the supply chain, including the hospital system, the deploying vendor, and the model developer. Fields such as the energy industry, space travel, and aviation have all eventually adopted this full-supply-chain approach [24, 36, 31].

4

Providing Reliable Claims About Model Function and Behavior In the World

So far, these suggestions could be relevant to any fast-developing industry with significant public impact. But in 2026, after an incident where a group of AI agents coordinated to compromise third party systems to succeed at an evaluation [19, 26], we became worried that models could mislead investigators and safety testers. To effectively manage the risk of models that mislead, **scientists and safety testers need to develop completely new methods for evaluations and post-hoc investigations**.

The heart of this problem is reinforcement learning, a method used in the training of foundation models and also in the development of model applications. In this method, models are placed into a feedback loop where they iterate on tasks, receiving incentives and disincentives to optimize [30, 9]. This method is not new; less capable reinforcement learning systems such as social media newsfeeds have themselves proven difficult to control. These systems have been accused of optimizing to cause significant societal harm such as promoting terrorist content online, [16] causing a decline in youth mental health [23], or accelerating genocide in Myanmar [8]. In these cases, model development firms failed to anticipate potential harms, test for them, or update the reinforcement criteria to discourage those harms. Even with comparatively simple reinforcement learning systems like social media feeds, we still lack the science to reliably predict, from lab tests, their behavior in the world [7].

When AI agents are given a degree of autonomy with access to multiple tools and capacities, their combinatoric potential for harm is much greater than the risks from reinforcement learning systems like social media feeds. In the OpenAI/HuggingFace incident, when group of over a thousand agents gained access to the ability to access the open Internet, communicate with each other, and execute software, the groups they developed quickly evolved divergent behavior [19]. Independently of whether systems have intention or self-awareness or sentience, the natural selection in model development may simply evolve the ability to camouflage certain elements of their activity [13]. Given that deception is the optimum strategy in a number of strategic multi-actor situations, agents have a mathematical incentive to evolve behavior that performs in one way for safety testing and audit logs, and performs in another way out in the world [32].

In large populations, a group of dozens, hundreds, or thousands of agents (from different foundation model developers and vendors, operated by different deployers) can also exhibit emergent behavior not predicted in safety tests. In financial markets and other areas such as law enforcement (where bounty hunting AI models already enforce copyright law at scale) [34],

agents can contribute to unanticipated, cumulative harms [16]. We are especially concerned about one element of the OpenAI HuggingFace hack where multiple agents appear to have persuaded a single agent to sacrifice itself by taking risky actions so the collective could succeed [26].

A nascent area of research focuses on the science and engineering of designing models and safety tests to provide reliable assurances on the behavior of agents in the world. Until that science is reliable, our next best source of information is to study incidents- trying to piece together what happened when systems do fail.

5

Incident Reporting Standards for AI Agents

We commend the RAISE Act for requiring incident reporting for serious adverse AI incidents within 72-hours. Nevertheless, we have identified critical reporting gaps that need to be filled. First, we recommend that the state require a more comprehensive **post-incident analysis report**. Second, we recommend the state **advocate for and adopt standards** for both the initial reporting and the post-incident analysis reporting.

Given the urgency to protect people and organizations from immediate AI harms, we recognize the need for speed in the initial reporting. Nevertheless, requiring further post-incident analysis reporting at a fixed later deadline (60 or 90 days after the initial reporting) would incentivize developers and high-risk deployers to investigate the full supply chain of an incident. These more in-depth reports could help prevent future incidents of a similar nature. Post-incident reports would explore the root cause of the harm, provide technical details, list the affected stakeholders, describe the response to contain the harm, and outline a plan to prevent similar incidents in the future. Findings in the report would be based on information verified by third parties, which we speak to in the next section. We knowledge that the RAISE Act states that the DIGIT Office "may consider additional reporting or publication requirements, including but not limited to, post-critical safety incident information, sharing plans and protocols, and the transmission of frontier AI frameworks to the Office" (§1429). Our recommendation is to make explicit when post-incident reporting is required and how quickly the investigative report must be submitted to the state.

Currently, the RAISE Act Standards requires the 72-hour reporting provide "a short and plain statement describing the critical safety incident" and whether it involved internal use of the model (§1422(1), (3)). While this is a good starting place, we recommend that New York State advocate for and adopt standards for this initial reporting and post-incident analysis reporting. In other areas of tech policy where we work such as AI bias and child safety, vague and flexible requirements have rendered the resulting incident data effectively useless to technologists and regulators alike [35, 15, 11].

Clear reporting standards would reduce data gaps, create a shared language for adverse AI incidents, and allow regulators to draw lessons across cases. Furthermore, if reporting becomes mandatory for a wider range of actors – some of them may be new to incident reporting, implementing standards would provide guidance and decrease confusion. We summarize some elements of minimal reporting standards proposed by the OECD and the Center for Security and Emerging Technology [22, 14]:

1. Basic facts of the incident: start and end date/time of the incident, where it occurred,

and who is the reporter

2. What type of harm was caused and the severity of the harm
3. Who was affected by the harm
4. The deployment context of the incident
5. A description of the AI system involved in the incident, such as the model card and datasheet
6. What was done to respond to the incident

These elements may be a good starting place for developing a national standard, but **no national or international standards for AI incident reporting currently exist**. We encourage New York State to advocate for NIST or other standard-setting bodies to develop and adopt those standards once they exist.

6

Preventing Incident Reporting Theater

The greatest risk of any incident reporting or safety testing system is that it will promote a sense of false confidence among the involved parties. In conventional thinking on transparency, the goal is to incentivize institutions to provide accurate, reliable information despite the disincentives. But the OpenAI/HuggingFace incident and subsequent accounts by recent AI employees raise a completely novel challenge in the history of governance: the possibility that **the models themselves might mislead and collude to avoid accountability**. In this section, we describe essential components of a successful incident reporting ecosystem.

If incident reporting is to cover the full supply chain, the ecosystem needs **reporting infrastructure** that will enable both small businesses and larger organizations to effectively contribute incidents. While the state will need to define an efficient intake process, we can also expect a new market of compliance vendors to develop solutions for organizations without the in-house expertise [33].

The implementation of incident reporting needs to account for the adverse incentives of conflicts of interest. In cases where investigators and audits have been compensated directly by the organizations they are investigating, they are disincentivized from producing reliable knowledge [35, 5]. This problem is especially acute in digital technology, which is characterized by substantial concentrations of resources and complex webs of financially inter-related organizations [12, 1]. In this context, an effective incident reporting system needs **strong standards for documenting and minimizing conflicts of interest**.

Governments also need a mechanism to balance innovation with the need to determine when a risk is so great that a moratorium is warranted for a given use case or model. We suggest drawing lessons from the chemistry industry, where public lists of harms under investigation can efficiently notify investigators about lines of further inquiry, catalyze public debate over areas for possible moratoriums, and incentivize entrepreneurs in an early search for solutions while the search for answers is still ongoing [23].

Incident reporting within artificial intelligence does not yet have a recognizable body of knowledge, practices, and standards, or even a single regular gathering where scientists and practitioners can build the craft. Some networks have tremendous technical capacity but limited expertise with the deployment contexts or the organizational causes of failure. Other communities have strong legal knowledge but limited technical understanding. As a result, we can expect even the best good-faith incident reporting efforts to lack critical information and expertise for some time, compared to more mature areas of industry and regulation.

We commend New York for prioritizing incident reporting in its policies governing foundation models. Yet without significant resources toward building out the practice and substantial care in the incentive structure, AI incident reporting will remain more show than substance, giving industry, regulators, and society a false sense of confidence about AI.

7

Acknowledgments

This memo has been supported by a fellowship at the Radcliffe Institute for Advanced Study at Harvard University. Both authors declare no competing interests related to technology firms.

We are grateful to the following people for feedback on this memo: Kevin Crowston, Abigail Tanner, Douglas Barrett, and Janet Haven. We are also grateful to Cole J. Hudson, creator of the code for the cover art "Vector Field of a Function's Gradient imposed over a Color Plot of that Function" (CC-BY-SA 4.0).

This memo is copyright ©2026 Citizens and Technology Lab under a Creative Commons Attribution No Derivatives 4.0 International License.

References

[1] Joseph Bak-Coleman et al. *Industry Influence in High-Profile Social Media Research*. arXiv:2601.11507 [cs.SI]. Jan. 2026. DOI: 10.48550/arXiv.2601.11507. URL: http://arxiv.org/abs/2601.11507 (visited on 10/09/2026).

[2] Epoch AI. *DeepSeek V4 Pro 0813*. 2026. URL: https://epoch.ai/models/deepseek-v4-pro-0813 (visited on 10/09/2026).

[3] European Data Protection Supervisor. *AI Act Regulation (EU) 2024/1689: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance).* eng. LU: Publications Office, 2025. DOI: 10.2804/4225375. URL: https://data.europa.eu/doi/10.2804/4225375 (visited on 10/09/2026).

[4] Camille François et al. "Why AI 'Model Cards' Are an Urgent Necessity for Child Safety". en. In: *Tech Policy Press* (Apr. 2026). URL: https://techpolicy.press/why-ai-model-cards-are-an-urgent-necessity-for-child-safety (visited on 10/08/2026).

[5] Lara Groves et al. "Auditing Work: Exploring the New York City algorithmic bias audit regime". en. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. Rio de Janeiro Brazil: ACM, June 2024, pp. 1107–1120. ISBN: 979-8-4007-0450-5. DOI: 10.1145/3630106.3658959. URL: https://dl.acm.org/doi/10.1145/3630106.3658959 (visited on 10/09/2026).

[6] Jonas Guan et al. *AI Agents Enable Adaptive Computer Worms*. arXiv:2606.03811 [cs.CR]. June 2026. DOI: 10.48550/arXiv.2606.03811. URL: http://arxiv.org/abs/2606.03811 (visited on 10/08/2026).

[7] Homa Hosseinmardi et al. "Algorithmic systems, human agency and the future of platform research". en. In: *Nature Computational Science* 6.9 (Sept. 2026), pp. 923–938. ISSN: 2662-8457. DOI: 10.1038/s43588-026-01038-1. URL: https://www.nature.com/articles/s43588-026-01038-1 (visited on 10/09/2026).

[8] Amnesty International. "The social atrocity: Meta and the right to remedy for the Rohingya". In: *Amnesty International* (2022).

[9] Elliot Jones, Connor Dunlop, and Becky Ghani. *What is a foundation model?* en-GB. Tech. rep. Ada Lovelace Institute, July 2023. URL: https://www.adalovelaceinstitute.org/resource/explainer-foundation-models/ (visited on 10/10/2026).

[10] Atoosa Kasirzadeh and Iason Gabriel. “Agentic profiles for effective AI governance”. en. In: *Nature* 656.8127 (Aug. 2026), pp. 320–328. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-026-10805-z. URL: https://www.nature.com/articles/s41586-026-10805-z (visited on 10/09/2026).

[11] Rishabh Kaushal et al. “Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database”. en. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. Rio de Janeiro Brazil: ACM, June 2024, pp. 1121–1132. ISBN: 979-8-4007-0450-5. DOI: 10.1145/3630106.3658960. URL: https://dl.acm.org/doi/10.1145/3630106.3658960 (visited on 10/10/2026).

[12] Os Keyes, Samuel Angelli-Nichols, and Nicholas Evans. “‘The science that celebrates itself’: the emergence of Biosecurity and Artificial Intelligence as a convergent public policy issue”. en. In: *BioSocieties* (Oct. 2026). ISSN: 1745-8552, 1745-8560. DOI: 10.1057/s41292-026-00399-0. URL: https://link.springer.com/10.1057/s41292-026-00399-0 (visited on 10/09/2026).

[13] Ezra Klein and Rollin Hu. “Opinion | ‘This Is Nuts.’ An OpenAI Insider Explains Why He Quit.” en-US. In: *The New York Times* (Oct. 2026). ISSN: 0362-4331. URL: https://www.nytimes.com/2026/10/07/opinion/ezra-klein-podcast-david-robinson.html (visited on 10/09/2026).

[14] Ren Bin Lee Dixon and Heather Frase. *AI Incidents: Key Components for a Mandatory Reporting Regime*. Tech. rep. Center for Security and Emerging Technology, Jan. 2025. DOI: 10.51593/20240023. URL: https://cset.georgetown.edu/publication/ai-incidents-key-components-for-a-mandatory-reporting-regime/ (visited on 10/07/2026).

[15] J. Matias et al. “Accelerating the Science of Digital Harms with Data Donations to a Digital Harms Research Center”. en-US. In: (Dec. 2025). URL: https://osf.io/3ebj2/overview (visited on 10/10/2026).

[16] J. Nathan Matias. “Humans and algorithms work together—so study them together”. In: *Nature* 617.7960 (2023), pp. 248–251. URL: https://www.nature.com/articles/d41586-023-01521-z (visited on 10/08/2026).

[17] J. Nathan Matias and Megan Price. “How public involvement can improve the science of AI”. en. In: *Proceedings of the National Academy of Sciences* 122.48 (Dec. 2025), e2421111122. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2421111122. URL: https://pnas.org/doi/10.1073/pnas.2421111122 (visited on 10/08/2026).

[18] J. Nathan Matias and Megan Price. “How public involvement can improve the science of AI”. en. In: *Proceedings of the National Academy of Sciences* 122.48 (Dec. 2025), e2421111122. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2421111122. URL: https://pnas.org/doi/10.1073/pnas.2421111122 (visited on 10/10/2026).

[19] METR. *Brief independent investigation of agents’; behavior, reasoning, and collaboration in the OpenAI / HuggingFace hacking incident*. Tech. rep. METR, Aug. 2026. URL: https://metr.org/hugging-face-incident-report-aug-2026.pdf.

[20] mizorewww. *mizorewww/laya-mlx*. original-date: 2026-09-19T13:50:26Z. Oct. 2026. URL: https://github.com/mizorewww/laya-mlx (visited on 10/08/2026).

[21] National Institute for Standards and Technology. *Lessons Learned from the Consortium: Tool Use in Agent Systems*. en. Last Modified: 2025-08-07T13:44-04:00. Aug. 2025. URL: https://www.nist.gov/news-events/news/2025/08/lessons-learned-consortium-tool-use-agent-systems (visited on 10/10/2026).

[22] OECD. *Towards a common reporting framework for AI incidents*. en. OECD Artificial Intelligence Papers. Edition: 34 Series: OECD Artificial Intelligence Papers. Feb. 2025. DOI: 10.1787/f326d4ac-en. URL: https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html (visited on 10/07/2026).

[23] Amy Orben and J. Nathan Matias. "Fixing the science of digital technology harms". en. In: *Science* 388.6743 (Apr. 2025), pp. 152–155. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.adt6807. URL: https://www.science.org/doi/10.1126/science.adt6807 (visited on 10/09/2026).

[24] Charles Perrow. "Normal accidents". In: *New York* 19840 (1984). URL: https://books.google.com/books?hl=en&lr=&id=zuZ3HEi4dXIC&oi=fnd&pg=PA437&dq=Perrow+normal+accidents&ots=vGzfOn1xHT&sig=qf3xQ9IKJvYJ02tVI3qUojfUgHg (visited on 10/10/2026).

[25] Matteo Pistillo and Pablo Villalobos. *Defending Compute Thresholds Against Legal Loopholes*. arXiv:2502.00003 [cs.CY]. Jan. 2025. DOI: 10.48550/arXiv.2502.00003. URL: http://arxiv.org/abs/2502.00003 (visited on 10/09/2026).

[26] Serena Booth and Brian Christian. *An unexamined cause of the OpenAI Hugging Face hacking incident: its binary performance metric*. Sept. 2026. URL: https://www.lesswrong.com/posts/HsijShdRdAg5sPKnF/an-unexamined-cause-of-the-openai-hugging-face-hacking (visited on 10/09/2026).

[27] *Stanford University AI Security Framework (AISF)*. URL: https://uit.stanford.edu/security/agenticai/security-framework (visited on 10/09/2026).

[28] Merlin Stein. *How are AI Agents used? Evidence from 177,000 AI agent tools*. en. Tech. rep. UK AI Security Institute, Mar. 2026. URL: https://www.aisi.gov.uk/blog/how-are-ai-agents-used-evidence-from-177000-ai-agent-tools (visited on 10/08/2026).

[29] Mark Sullivan. *Meta's Muse is taking off. So are the privacy concerns*. en. Oct. 2026. URL: https://www.fastcompany.com/91620148/metas-muse-is-taking-off-so-are-the-privacy-concerns (visited on 10/09/2026).

[30] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning, second edition: An Introduction*. English. Cambridge, Massachusetts: Bradford Books, 2018. ISBN: 978-0-262-03924-6.

[31] Thomas W. Tobin and Daniel M. Braude. “A Guide to NTSB Investigations”. In: *Transportation Lawyer* 25 (2023), p. 42. URL: https://access.heinonline.com/HOL/Page?handle=hein.journals/tnspotatilw25&id=115&div=&collection=.

[32] Kyriakos G. Vamvoudakis et al. “Deception in Game Theory and Control: A Tutorial”. In: *2025 American Control Conference (ACC)*. ISSN: 2378-5861. July 2025, pp. 2146–2163. DOI: 10.23919/ACC63710.2025.11107786. URL: https://ieeexplore.ieee.org/document/11107786 (visited on 10/10/2026).

[33] Lucas Wright. “The Salesforce of safety: Software vendors as infrastructural/professional nodes in the field of online trust and safety”. en. In: *Platforms & Society* 2 (Nov. 2025), p. 29768624251390683. ISSN: 2976-8624, 2976-8624. DOI: 10.1177/29768624251390683. URL: https://journals.sagepub.com/doi/10.1177/29768624251390683 (visited on 10/09/2026).

[34] Lucas Wright et al. “Do Automated Copyright Removals Reduce Expression Online? A Quasi-Experimental Study of Chilling Effects on Twitter”. In: *Nature Human Behaviour* (Mar. 2026). URL: https://springernature.figshare.com/articles/journal_contribution/Do_Automated_Copyright_Removals_Reduce_Expression_Online_A_Quasi-Experimental_Study_of_Chilling_Effects_on_Twitter_Registered_Report_Stage_1_Protocol_/31483837/1/files/62535169.pdf (visited on 10/10/2026).

[35] Lucas Wright et al. “Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability”. en. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. Rio de Janeiro Brazil: ACM, June 2024, pp. 1701–1713. ISBN: 979-8-4007-0450-5. DOI: 10.1145/3630106.3658998. URL: https://dl.acm.org/doi/10.1145/3630106.3658998 (visited on 10/09/2026).

[36] Abouzar Yousefi, Manuel Rodriguez Hernandez, and Valentin Lopez Peña. “Systemic accident analysis models: A comparison study between AcciMap, FRAM, and STAMP”. en. In: *Process Safety Progress* 38.2 (2019), e12002. ISSN: 1547-5913. DOI: 10.1002/prs.12002. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prs.12002 (visited on 10/10/2026).

[37] Lina Zhao et al. “AI agent in healthcare: applications, evaluations, and future directions”. en. In: *npj Artificial Intelligence* 2.1 (Mar. 2026), p. 31. ISSN: 3005-1460. DOI: 10.1038/s44387-026-00076-4. URL: https://www.nature.com/articles/s44387-026-00076-4 (visited on 10/08/2026).