RECOMMENDATION: Improve Monitoring of Emerging Risks from AI through Adverse Event Reporting

The National Artificial Intelligence Advisory Committee (NAIAC)

November 2023

RECOMMENDATIONS

Recommendation:
Pilot an adverse event reporting system for AI.

NAIAC recommends piloting an adverse event reporting system that would allow developers, deployers, and users to report harmful post-deployment events stemming from AI systems.

The reporting system could be two-tiered. First, reporting could be mandated for a limited set of emerging risks that are of particular high concern, such as those posing national security threats (e.g., biorisk) and known, actual instances of serious injury, substantial damage, or death. Alternatively, reporting of adverse events (see Reportable Events section below) could be mandatory if arising from the use of safety-impacting and rights-impacting AI, as defined in the Office of Management and Budget's draft policy on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence.[1]

Reporting should be focused around the most concrete and severe risks arising directly from the use of AI, as described in the section below under "reportable events," rather than claims for which assessment would be more of a matter of evidentiary debate. Second, voluntary reporting could be instituted for other serious harms. Reported events could then be referred to agencies with existing regulatory authorities (e.g., Food and Drug Administration, Consumer Financial Protection Bureau, Bureau of Industry and Security of the Department of Commerce[2]) or expose

[1] Safety-Impacting AI is defined as "AI that has the potential to meaningfully impact the safety of: 1. Human life or well-being, including loss of life, serious injury, bodily harm, biological or chemical harms, occupational hazards, harassment or abuse, or mental health, including both individual and community aspects of these harms; 2. Climate or environment, including irreversible or significant environmental damage; 3. Critical infrastructure, including the critical infrastructure sectors defined in Presidential Policy Directive 2143 and the infrastructure for voting and protecting the integrity of elections; or, 4. Strategic assets or resources, including high-value property, information marked as sensitive or classified by the Federal Government, and intellectual property." Rights-Impacting AI is defined as "AI whose output serves as a basis for decision or action that has a legal, material, or similarly significant effect on an individual's or community's: 1. Civil rights, civil liberties, or privacy, including but not limited to freedom of speech, voting, human autonomy, and protections from discrimination, excessive punishment, and unlawful surveillance; 2. Equal opportunities, including equitable access to education, housing, credit, employment, and other programs where civil rights and equal opportunity protections apply; or 3. Access to critical resources or services, including healthcare, financial services, social services, transportation, non-deceptive information about goods and services, and government benefits or privileges." Draft Policy, Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, Office of Management and Budget, https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf.

[2] The Bureau of Industry and Security (BIS) regulates the export of commercial goods, including dual use technologies that require licenses to export. Dual Use Export Licenses, BIS,

gaps in existing regulatory authority. The reports would be kept confidential and for use by regulators and would provide a concrete evidence base for other potential actions, including the formulation of actionable response frameworks.[3]

An adverse event reporting system addresses the core informational challenge of AI regulation, namely an improved understanding of high risks associated with AI. This information would be of critical importance for institutions that respond to AI incidents, such as a possible AI Lead Rapid Response Team (ALRT) as previously recommended by NAIAC.

In addition, such a policy has several advantages. First, an adverse event reporting system is easier to administer and does not require extensive technical or institutional capacity to implement. Second, such a policy would aggregate information, reducing information asymmetries between regulators and industry. As a result, policymakers will have more complete information about the need, if any, for further regulation and resources needed to address harms identified through adverse event reporting. Third, reporting would directly provide information about risks, and not be contingent on arbitrary thresholds such as model size or computing resources used for a model. For example, a leading paper documenting dual use of AI to discover toxic chemical compounds involved a smaller-scale model that would not be covered by regulatory proposals contingent on model size or certain capabilities.[4] Finally, adverse event reports would capture dynamic and evolving risks more effectively than a one-time registration requirement, or other static test and evaluation approaches.

---

https://www.bis.doc.gov/index.php/all-articles/2-uncategorized/91-dual-use-export-licenses#:~:text=Dual%20use%20export%20licenses%20are,crime%20control%2C%20or%20terrorist%20concerns. The Department of State implements the International Traffic in Arms Regulation (ITAR) and its United States Munitions List, which implements the Arms Export Control Act. Directorate of Defense Trade Controls, Dep't of State, https://www.state.gov/bureaus-offices/under-secretary-for-arms-control-and-international-security-affairs/bureau-of-political-military-affairs/directorate-of-defense-trade-controls-pm-ddtc/. The Secretary of the Treasury chairs the Committee on Foreign Investment Screening (CFIUS), which reviews foreign investment transactions into and out of the United States for national security risks. *The Committee on Foreign Investment in the United States (CFIUS)*, Dep't of the Treasury, https://home.treasury.gov/policy-issues/international/the-committee-on-foreign-investment-in-the-united-states-cfius; *Outbound Investment Program*, Dep't of the Treasury, https://home.treasury.gov/policy-issues/international/outbound-investment-program.

[3] The adverse event catalog could provide the basis for responsive actions by the government. For examples of how an adverse event catalog could inform other government action see the National AI Advisory Committee recommendations on "Creating Institutional Structures to Support Safer AI Systems," https://ai.gov/wp-content/uploads/2023/11/Recommendations_Creating-Institutional-Structures-to-Support-Safer-AI-Systems.pdf.

[4] Fabio Urbina et al., "Dual Use of Artificial Intelligence-powered Drug Discovery," *Nature machine intelligence* vol. 4, 3 (2022): 189-191, https://www.nature.com/articles/s42256-022-00465-9.

Policy Design: While we recommend piloting with a limited set of national security risks (e.g., biorisks) and known, actual instances of serious injury, substantial damage, or death, an adverse event reporting system could ultimately be adapted to capture a wide range of risks.[5] In particular, the following features of a reporting system may be flexibly designed and tailored based on the pilot to address salient risks and provide more information about AI developments.[6]

Reportable Events: Policymakers can define what events are reportable and what information must be shared. We recommend that the focus should be on events for which AI plausibly played a substantive role, rather than including everything negative involving a system with AI. It may not always be clear whether the AI components were important causes, and so additional information may be required. This approach will thus require balancing the government's desire for more complete information against concerns of evasion or undue burden. In particular, if reporting requirements are too onerous, then companies may fail to promptly comply. Entities could report both realized adverse events as well as near misses,[7] similar to the Federal Aviation Administration[8] mandatory near-miss reporting protocols and Occupational Safety and Health Administration[9] suggested near-miss reporting policy for employers. Vulnerabilities that could result in adverse events could also be reported, similarly to common practice in cybersecurity.[10]

Regardless of the exact specifications of reportable events, we recommend first piloting the system with emphasis on the most acute risks and adverse events, such as national security concerns and known, actual instances of serious injury,

[5] "Rationales, Mechanisms, and Challenges to Regulating AI: A Concise Guide and Explanation," NAIAC, July 2023, https://www.ai.gov/wp-content/uploads/2023/07/Rationales-Mechanisms-Challenges-Regulating-AI-NAIAC-Non-Decisional.pdf.

[6] "National Artificial Intelligence Advisory Committee Year 1 Report," NAIAC, May 2023, https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf.

[7] The AI Incident Database defines "nearly harmed" as "a chain of events that easily could have caused harm, but some external factor kept the harm from occurring" where the AI system, although not necessarily the only factor, is the "but-for"cause (i.e., "that is, if the AI system hadn't acted in the way it did, there would have been no significant chance that the harm would occur"). "Editor's Guide," AI Incident Database, https://incidentdatabase.ai/editors-guide/.

[8] For instance, the Near Midair Collision Reporting System. "ENR 1.16 Safety, Hazard, and Accident Reports," Federal Aviation Administration, https://www.faa.gov/air_traffic/publications/atpubs/aip_html/part2_enr_section_1.16.html.

[9] "Near Miss Reporting Policy," Occupational Safety and Health Administration, https://www.osha.gov/sites/default/files/2021-07/Template%20for%20Near%20Miss%20Reporting%20Policy.pdf.

[10] For example, the AI Vulnerability Database is a non-governmental, open-source effort to document and identify "failure modes" for AI and "[b]uild out a functional taxonomy of potential AI harms across the coordinates of security, ethics, and performance." The database has an "effect view" focused on *risks* of the AI system and a "lifecycle view" to identify risks across six stages of an AI model's development and deployment. AI Vulnerability Database, https://avidml.org/; "Introduction," AI Vulnerability Database, https://avidml.gitbook.io/doc/taxonomy/introduction.

substantial damage, or death. This initial focus could also reveal potential misuses or abuses of the adverse event reporting system since the data would naturally receive extra scrutiny. Definitions of adverse events, incidents, and national security risks used by other departments and agencies[11] can inform the initial focus of the AI

[11] For example, the Food & Drug Administration defines a "serious adverse event" if it "results in any of the following outcomes: Death, a life-threatening adverse event, inpatient hospitalization or prolongation of existing hospitalization, a persistent or significant incapacity or substantial disruption of the ability to conduct normal life functions, or a congenital anomaly/birth defect." "21 CFR 312.32, CFR," U.S. Food & Drug Administration, https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=312.32#:~:text=An%20adverse%20event%20%20or%20suspected%20adverse%20reaction%20is%20considered%20%22serious,hospitalization%2C%20a%20pe%20rsistent%20or%20significant; "What is a Serious Adverse Event?" U.S. Food & Drug Administration, https://www.fda.gov/safety/reporting-serious-problems-fda/what-serious-adverse-event; An "aircraft accident" for National Transportation Safety Board reporting is one "in which any person suffers death or serious injury, or in which the aircraft receives substantial damage," with "serious injury" defined as injuries resulting in extended hospitalization or specified injuries (e.e., bone fractures, hemorrhages, second- or third-degree burns over more than 5 percent of body) and "substantial damage" meaning "damage or failure which adversely affects the structural strength, performance, or flight characteristics of the aircraft, and which would normally require major repair or replacement of the affected component . . . ." "49 CFR 830.2," Cornell Law School, Legal Information Institute, https://www.law.cornell.edu/cfr/text/49/830.2; For national security risks, the Commission on National Interests provides the canonical delineation of the five "vital national interests" that are "the conditions that are strictly necessary to safeguard and enhance Americans' survival and well-being in a free and secure nation" ("to prevent the threat of an attack of weapons of mass destruction on U.S. soil or its military abroad; to ensure U.S. allies' survival and cooperation to shape an international system in which we can thrive; to prevent the emergence of hostile powers on U.S. borders; to ensure the viability of major global systems; and to establish productive relations with nations that could become adversaries"). "America's National Interests," The Commission on America's National Interests, July 2000, https://www.belfercenter.org/sites/default/files/legacy/files/amernatinter.pdf; The 2022 National Security Strategy also delineates vital and extremely important national interests. "National Security Strategy," White House, 2022, https://www.whitehouse.gov/wp-content/uploads/2022/10/Biden-Harris-Administrations-National-Security-Strategy-10.2022.pdf; The factors that the Committee on Foreign Investment in the United States (CFIUS) considers when evaluating the risk a foreign investment poses to U.S. national security, including the factors highlighted in the 2022 Executive Order on Ensuring Robust Consideration of Evolving National Security Risks by the Committee on Foreign Investment in the United States, can also inform how an AI adverse reporting system should consider national security risks. "The Committee on Foreign Investment in the United States (CFIUS)," U.S. Department of Treasury, https://home.treasury.gov/policy-issues/international/the-committee-on-foreign-investment-in-the-united-states-cfius; "CFIUS Executive Order on Evolving National Security Risks and CFIUS Enforcement Guidelines," Congressional Research Service, May 26, 2023, https://crsreports.congress.gov/product/pdf/IF/IF12415.

adverse event reporting system[12], with the definition evolving as more information and trends are identified through the reporting.

Reporting Entities: The system can tailor who is eligible or subject to report: individuals, companies, developers, users. The design and incentives of the event reporting mechanism will need to be thoughtfully structured to ensure that entities do not have incentives to shift away from internal testing both prior to and after deployment (thereby shifting the burden onto users or the public) merely in order to evade reporting requirements. For example, adverse event reporting during pre-market internal tests (e.g., red teaming) may be desirable where developers are already required to engage in certain pre-market testing, but voluntary reporting may be better suited to in some scenarios to avoid disincentivizing internal testing.

Mandatory and Voluntary Reporting Standards: The system can be designed so that mandatory reporting is limited to high risk events that pose substantial national security risk (e.g., bioweapons risk) or and resulted in known, actual instances of serious injury, substantial damage, or death — drawing upon definitions used in other adverse event reporting regimes.[13] Voluntary reporting could include a wider range of potential harms to enable improved monitoring of emerging risks. This would mirror the FDA Adverse Event Reporting System (FAERS), which allows anyone to submit reports, including patients, healthcare providers, and manufacturers.[14]

---

[12] Efforts are underway to define acute risks posed by AI. For example, Anthropic's AI Safety Levels define highest risk ASL-4 systems as involving "critical catastrophic misuse risk" (i.e., the AI system is the "primary source of national security risk in a major area (such as cyberattacks or biological weapons), rather than just being a significant contributor"), "autonomous replication in the real world" (i.e., "unambiguously capable of replicating, accumulating resources, and avoiding being shut down in the real world indefinitely . . . "), or involving "autonomous AI research" (i.e., the "weights [in the model] would be a massive boost to a malicious AI development program"). However, Anthropic notes that the exact risks are not yet known. "Anthropic's Responsible Scaling Policy," Anthropic, September 19, 2023: 14, https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf; The paper "Frontier AI Regulation: Managing Emerging Risks to Public Safety" suggests anchoring on the yet-to-be-defined "sufficiently dangerous capabilities," which could refer to a capability that poses a "severe risk to public safety." Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," OpenAI, July 6, 2023, https://openai.com/research/frontier-ai-regulation; The AI Incident Database notes it has an "adaptive" understanding of incidents that leverages an evolving "Editor's Guide" that defines terms such as "AI incident" and "real world harm." "Defining an 'AI Incident,'" AI Incident Database, https://incidentdatabase.ai/research/1-criteria/; "Editor's Guide," AI Incident Database, https://incidentdatabase.ai/editors-guide/; The AI Vulnerability Database has, and is continuously refining, a taxonomy of security-related vulnerabilities, https://avidml.gitbook.io/doc/taxonomy/introduction, and MITRE Atlas details different AI vulnerabilities that can be intentionally exploited, https://atlas.mitre.org/.

[13] See, e.g., footnotes 6-11.

[14] "Questions and Answers on FDA's Adverse Event Reporting System (FAERS)," U.S. Food & Drug Administration, https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers.

Advantages and Complementarities to Adverse Event Reporting

Adverse event reporting provides complementary capabilities relative to other proposed regulations, such as registration and licensing of AI systems or practitioners.[15] Licensing procedures may require that systems, companies, or practitioners meet certain criteria (e.g., training, education, safeguards) before developing or deploying AI technology.[16] Registration requirements may involve mandated disclosure of training data, model attributes, or capabilities of AI systems and restrict use of unregistered models. AI licensing proposals largely address risks through pre-market approval, while AI registration facilitates monitoring to address post-market risk.

In contrast, adverse event reporting improves government information about potential harms of AI systems by capturing diverse realized risks. Adverse event reporting also would provide an independent basis for government evaluation of risks and harms to inform future action, without relying on only selective information. In addition, adverse event reporting does not create potential barriers to entry, maintaining a level playing field as the AI ecosystem matures. Finally, adverse event reporting places burdens primarily on developers and deployers of AI systems that pose greater risks than AI systems resulting in less frequent reportable events.

Adverse event reporting helps to remedy significant information gaps about emergent risks of AI systems, and so can be part of a group of regulatory proposals. Model registration, for instance, can enable adverse event reporting tied to specific models. Registration and adverse event reporting could work in parallel to monitor both AI activity and emerging risks.[17]

[15] Neel Guha, Christie M. Lawrence, et al. "AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing," *George Washington University Law Review* (2024).
[16] NAIAC, "Rationales, Mechanisms, and Challenges to Regulating AI: A Concise Guide and Explanation."
[17] In the securities context, there is some evidence that registration decreases misreporting. Colleen Hongsberg, "Hedge Fund Regulation and Fund Governance: Evidence on the Effects of Mandatory Disclosure Rules," *Journal of Accounting Research*, vol. 57, 4 (September 2019): 845, https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-679X.12270 ("Misreporting decreased at the funds that were required to register with the SEC, and increased at the funds that withdrew from registration after the courts vacated the SEC's rule (although this result should be interpreted as descriptive because the decision to withdraw is highly endogenous).").

CONTEXT

A complete and accurate assessment of AI risks is essential to safeguard U.S. security, economic, and democratic interests. Yet the emerging risks of advanced AI models are not well understood,[18] posing significant challenges to regulation aimed at mitigating potential harms. Two major impediments to government action are the uncertainty surrounding potential risks associated with public access to, and deployment of, AI models, along with the rapid evolution of those risks as uses and contexts shift. Therefore, a natural first step for policy is to improve government capacity to identify, monitor, assess, and understand emerging risks and harms. AI governance and other responses require an understanding of harms and benefits, both potential and actual, but such understanding is currently often absent.

In other policy areas, adverse event reporting has been effectively used[19] to identify novel risks, monitor for emergent risks, identify trends, develop safety recommendations, render assistance to prevent future adverse events, and measure progress towards remediating risks. For instance, the Cybersecurity and Infrastructure Security Agency (CISA) has mandatory and voluntary reporting about known incidents,[20] known phishing attempts, malware, and vulnerabilities and also allows constituents and partners to share cyber threat indicators and defensive actions.[21] CISA's Vulnerability Disclosure Policy (VDP) Platform alone has enabled

[18] Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv (July 2023), https://arxiv.org/abs/2307.03718; "Anthropic's Responsible Scaling Policy," Anthropic, 2023, https://www-files.anthropic.com/production/files/responsible-scaling-policy-1.0.pdf.
[19] For example, NHTSA developed passenger side airbag performance requirements based on actual accident data: https://www.ntsb.gov/safety/safety-studies/Documents/SR0601.pdf.
[20] For example, the Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA) requires has mandatory reporting requirements, including that "covered entities" report to CISA "any covered cyber incidents within 72 hours after the entity reasonably believes the covered cyber incident occurred": https://www.cisa.gov/sites/default/files/2023-09/2023-summit-circia-508.pdf; 6 U.S.C. 681-681g.
[21] "Report to CISA," Cybersecurity & Infrastructure Security Agency, https://www.cisa.gov/report.

agencies to address more than a thousand bugs, or 84% of reported bugs.[22] Many systems aggregate information about product defects (e.g., Consumer Product Safety Commission, MedWatch) through mandatory and voluntary reporting.[23] Similarly, public health monitoring systems consist of mandated reporting by laboratories (e.g., for listed pathogens such as smallpox) and may include active disease investigation (e.g., outbreak investigations).[24] Or consider the FDA-TRACK system that conducts postmarket monitoring for drugs and other pharmaceutical agents once they have been approved by the FDA.[25] Mandatory reporting is often limited to incidents and adverse events actually known by regulated entities,

[22] Justin Doubleday, "CISA platform helps agencies uncover more than 1,000 cyber vulnerabilities," Federal News Network, August 25, 2023, https://federalnewsnetwork.com/cybersecurity/2023/08/cisa-platform-helps-agencies-uncover-more-than-1000-cyber-vulnerabilities/; "Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA), Cybersecurity & Infrastructure Security Agency, 2023, https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing/cyber-incident-reporting-critical-infrastructure-act-2022-circia#:~:text=Cyber%20Incident%20Reporting%20Requirements%3A%20CIRCIA,reasonably%20believes%20the%20incident%20occurred. ("When information about cyber incidents is shared quickly, CISA can use this information to render assistance and provide warning to prevent other organizations from falling victim to a similar incident. This information is also critical to identifying trends that can help efforts to protect the homeland."); Edward Graham, "Cyber incident reports will be shared with the agency under the soon-to-be implemented requirements of the Cyber Incident Reporting for Critical Infrastructure Act," NextGov/FCW, March 28, 2023, https://www.nextgov.com/cybersecurity/2023/03/new-cyber-reports-will-show-value-cisa-budget-investments-director-says/384540/ ("We can say, 'here are the number of critical incidents that occurred across our critical infrastructure this year,' and then we can measure the reduction given all of the improvements that we've put in place," Easterly added. "So we are on our journey to be able to give you very quantifiable metrics to allow us to articulate that return on investment.")

[23] The Consumer Product Safety Act mandates companies report to the CPSC certain harms caused by consumer products manufactured, imported, distributed, and/or sold by the company. However, consumers may voluntarily report safety problems. "Duty to Report to CPSC: Rights and Responsibilities of Businesses," Consumer Product Safety Commission, https://www.cpsc.gov/Business--Manufacturing/Recall-Guidance/Duty-to-Report-to-the-CPSC-Your-Rights-and-Responsibilities; "Who We Are - What We Do for You," U.S. Consumer Product Safety Commission, https://www.cpsc.gov/Safety-Education/Safety-Guides/General-Information/Who-We-Are---What-We-Do-for-You#: ~:text=If%20you've%20had%20a,the%20hearing%20and%20speech%20impaired. MedWatch include voluntary reporting by health professionals, patients, and consumers, with mandatory reporting for industry (e.g., with Form FDA 3500A). "Reporting Serious Problems to FDA," U.S. Food & Drug Administration, https://www.fda.gov/safety/medwatch-fda-safety-information-and-adverse-event-reporting-program/reporting-serious-problems-fda; "MedWatch Forms for FDA Safety Reporting," U.S. Food & Drug Administration, https://www.fda.gov/safety/medical-product-safety-information/medwatch-forms-fda-safety-reporting

[24] "Introduction to Public Health Surveillance," Public Health 101 Series, Centers for Disease Control and Prevention, 2014, https://www.cdc.gov/training/publichealth101/surveillance.html; "Specimen Collection and Transport Guidelines for Suspect Smallpox Cases," Centers for Disease Control and Prevention, https://www.cdc.gov/smallpox/lab-personnel/specimen-collection/specimen-collection-transport.html; "Epidemiological Investigation," Centers for Disease Control and Prevention, https://www.cdc.gov/smallpox/bioterrorism-response-planning/public-health/epidemiological-investigation.html.

[25] "FDA-TRACK: Center for Drug Evaluation & Research - Post-Approval Safety Monitoring," U.S. Food & Drug Administration, https://www.fda.gov/about-fda/fda-track-agency-wide-program-performance/fda-track-center-drug-evaluation-research-post-approval-safety-monitoring.

including through consumer complaints. Each of these uses constitutes an attempt to learn about unknown or unquantifiable risks and harms in the presence of substantial uncertainty, highlighting the potential benefit of adverse event reporting for AI.[26]

ABOUT NAIAC

The National Artificial Intelligence Advisory Committee (NAIAC) advises the President and the White House National AI Initiative Office (NAIIO) on the intersection of AI and innovation, competition, societal issues, the economy, law, international relations, and other areas that can and will be impacted by AI in the near and long term. Their work guides the U.S. government in leveraging AI in a uniquely American way — one that prioritizes democratic values and civil liberties, while also increasing opportunity.

NAIAC was established in April 2022 by the William M. (Mac) Thornberry National Defense Authorization Act. It first convened in May 2022. It consists of leading experts in AI across a wide range of domains, from industry to academia to civil society.

https://www.ai.gov/naiac/

###

[26] Current attempts to aggregate information about AI harms include the AI Incident Database, https://incidentdatabase.ai.