Establishing an AI Incident Reporting Database

Policy Memo by David Robusto, Americans for Responsible Innovation; with support from Inclusive Abundance

As artificial intelligence (AI) rapidly advances, it holds the potential to transform healthcare, education, economic development, and more through increased productivity and personalization. However, without careful oversight, this promising technology can experience harmful failures and be subject to dangerous misuse. To mitigate these risks, policymakers need to have clear insight into real-world issues affecting AI systems. Congress has the opportunity to collect critical data on AI failures and harms by establishing a voluntary national AI Incident Reporting Hub (AIIRH).

The AIIRH would be a resource for developers, whistleblowers, researchers, and those negatively affected by AI systems to identify and share information with the federal government about AI failures, accidents, security breaches, and other potentially hazardous incidents. The AIIRH would aggregate these incidents through a voluntary reporting system targeting harms like discrimination, cyber threats, AI-enabled NCII, and CBRN risks. Such a system would provide the government with crucial, high-quality data on this rapidly changing threat landscape and allow policymakers to create guidelines and ground future regulatory and legislative interventions in real-world incidents.

In order to provide the public sector with maximal data the AIIRH should cover a wide range of incidents. To this end, the system would (at least initially) utilize the relevant OECD definition (*Figure 1)* and also accommodate "AI hazards" which are any situations that "*could plausibly lead to an AI incident."*

OECD AI INCIDENT DEFINITION

*An **AI incident** is an event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to any of the following harms:*

***(a)** injury or harm to the health of a person or groups of people;*
***(b)** disruption of the management and operation of critical infrastructure;*
***(c)** violations of human rights or a breach of obligations under the applicable law intended to protect fundamental, labour and intellectual property rights;*
***(d)** harm to property, communities or the environment.*

FIGURE 1

With this scope, the AIIRH would cover a wide range of AI harms, from discriminatory (e.g., a company providing a tenant-screening algorithm was accused of discriminating against certain Black and Hispanic applicants, subsequently agreeing to pay over $2 million and make

AMERICANS for RESPONSIBLE INNOVATION

Inclusive Abundance

changes to their system, to financial (e.g., automated trading systems or AI agents crashing markets), to catastrophic (e.g., the misuse of AI to advance chemical, biological, radiological, and nuclear (CBRN) threats).

Establishing this system through legislation would instill confidence in its longevity and security, as it would also allow strict cybersecurity and confidentiality guarantees on reported information. Existing AI incident reporting systems that lack these legally-sanctioned guarantees serve more as repositories for incidents that receive media attention.

To preserve confidentiality, reporting forms should allow maximum information but *require* as little as possible, encouraging industry reporting without fear of leaking sensitive information and lowering the implied transaction costs of reporting. Contributors should also have the option to reveal their identity only to AIIRH staff and otherwise maintain anonymity.

To this end, the system would be modeled on similar voluntary incident reporting systems managed and funded by the federal government. These voluntary and confidential systems have achieved success in domains like cybersecurity (e.g., the National Vulnerability Database) and aviation safety (e.g., the confidential Aviation Safety Reporting System) and contributed to a sectoral culture of collaborating on risks. The AIIRH would also take inspiration from AI incident reporting systems run by groups like the Responsible AI Collaborative (RAIC)'s database.

An additional benefit of the AIIRH being a government-controlled system is that expert researchers, such as those at the National Institute of Standards and Technology (NIST), can analyze reported incidents to provide a basis for the development of future standards and guidance. One way they might do this is by building a taxonomy of different types of incidents.

NIST's AI expertise is one of several reasons why it is the natural choice to house and manage the AIIRH. Additionally, NIST has experience with incident reporting through running and conducting analysis on the National Vulnerability Database. Finally, NIST is a non-regulatory agency with excellent industry relationships, demonstrated by their AI safety consortium consisting of nearly 300 leading companies, helping to increase reporters' confidence in the AIIRH and incentivize reporting.

While NIST might be a trusted party to manage a confidential system, employees reporting credible dangers to the AIIRH should have additional guarantees against retaliation from their current or former employers in the form of whistleblower or "right to warn" protections. These protections are particularly relevant given growing concerns that top AI companies are preventing their current and former employees from speaking out on safety issues.

It is promising that the idea of an AI incident reporting system is gaining traction in Congress and Members have begun to introduce legislation on the topic. First, in May 2024, Senators Mark Warner (D-VA) and Thom Tillis (R-NC) introduced the Secure AI Act, which would

establish a public database for AI safety and security incidents. More recently, a House companion to the Secure AI Act was introduced and on September 25th, the AI Incident Reporting and Security Enhancement Act – which would mandate a report on establishing an incident database–passed out of the House Committee on Science, Space, and Technology. Americans for Responsible Innovation supports both of these efforts to tackle this crucial issue.

Overall, an AI incident reporting system would enable informed policymaking as the risks of AI continue to develop. By facilitating voluntary reports on serious AI risks and harms – such as CBRN risk, illegal discrimination, and cyber threats – the AIIRH would enable the U.S. government to collect and analyze high-quality data and, as needed, promulgate standards to reduce harms and the proliferation of dangerous capabilities. Incentivizing voluntary reporting can help preserve innovative and high-value uses of AI for society and the economy, while keeping policymakers up-to-date with the quickly evolving frontier in cases where regulatory oversight is paramount. As AI systems become more capable, the associated risks continue to grow, and the importance of this system will grow with them.