Category 1: Operational Safety & Reliability

Risks related to AI systems failing to operate safely, pursuing misaligned goals, or lacking robustness.

Note: This table is meant to provide illustrative examples and is not intended to cover all types of risks nor every event included under each type.

Type of Risk	Incidents(s) / Hazard(s)	Incident Response Analysis and Recommendations
Unsafe treatment of self-harm topics	• AI chat provider allegedly failed to escalate or intervene in teen conversation regarding self-harm (Nov 2023)	• Currently, the alleged role of AI systems in these incidents are only captured ad-hoc by media reporting • **CAISI:** Establish incident reporting standards and data processing guidance that allows incident reports from both consumers, media, and industry to be aggregated in order to understand the scale and scope of harms and identify any common cause or vector.
Unfair Decisions	• Alleged unjustified denial of Health Insurance Claims (Oct 2023) • Alleged denial of claims due to inaccurate algorithm (July 2023)	
Human Deception & Misalignment	• Anthropic Research discovers Alignment Faking (Dec 2024) • Apollo Research discovers AI "scheming" (Dec 2024) • Grok produces hateful content after update, contrary to provided instructions (July 2025)	• Currently, these findings and events have not warranted significant incident responses, but they are warning signs of larger deception and misalignment that could occur • **CAISI:** Establish guidance and best practices for industry and academia on responsible AI risk and incident disclosure during testing, development, and deployment. Note these could be either safety or security issues which warrant different approaches to the recommended timing and detail of disclosures. • **The AI-ISAC**, also proposed in the AI Action Plan, could serve as a hub for this information.
Loss of Control	• Customer report describes an AI coding tool making changes and deleting data, violating direct instructions. (July 2025) • AI Agent orders grocery delivery despite instructions and guardrails to prevent purchases without human intervention (Feb 2025)	• **CAISI:** Guidance on responding to incidents when AI agents taking undesired actions should include: when and how to shut down agents, identifying all undesired actions, and improving testing of the model during development and deployment. ◦ e.g. in the data deletion example, lacking a robust shutdown mechanism, the operators were unable to stop the tool by issuing direct instructions. • **The AI-ISAC**, also proposed in the AI Action Plan, could serve as a hub for this information.

Category 2: Security and Integrity

Risks related to weaknesses or flaws in AI systems themselves that allow unauthorized access to sensitive information contained in the system, unwanted visibility into the system, or improper control of the system.

Note: This table is meant to provide illustrative examples and is not intended to cover all types of risks nor every event included under each type.

Type of Risk	Incidents(s) / Hazard(s)	Incident Response Analysis and Recommendations
Data Privacy	• Privately shared AI chats hosted in publicly accessible locations that were later stored and archived (Aug 2025) • AI Agent provider exposed sensitive medical data online (Sept 2024)	• Currently, despite the lack of a comprehensive federal privacy law, sector-specific laws such as HIPAA, as well as state privacy and data breach laws, govern the response to data theft or leaks where applicable. • **CAISI:** Ensure federal data breach response guidance is suitable for data theft or leaks involving AI systems across the value chain (developers, deployers…) • **CAISI:** Ensure HIPAA data breach [requirements/best practices] are suitable for AI systems, including potential data leaks from AI systems trained on sensitive data.
Model Theft	• Meta's LLaMA leaks online before public release (March 2023) • Mistral leaks older model (Feb 2024)	• Currently, there is voluntary reporting of cybercrime and intellectual property crime • **CAISI:** Develop guidance for information sharing between industry and government, in coordination with the AI-ISAC, on specific threats to AI systems' security. • **CAISI:** Establish interagency collaboration and information sharing mechanisms in the case of theft of a powerful model that creates immediate cross-sector or national security concerns.
Prompt Injection	• Student obtains confidential information about an LLM directly from the LLM to demonstrate Prompt Injection attacks (Feb 2023)	• Currently, findings demonstrating how attackers could take control of AI systems have not materialized in severe incidents, but they are warning signs about weaknesses that could be exploited at a larger scale and become major security incidents. • **CAISI:** Establish guidance and best practices for industry and academia on responsible AI risk and incident disclosure during testing, development, and deployment. ◦ The **AI-ISAC**, also proposed in the AI Action Plan, could serve as a hub for this information.

Category 3: Abuse & Offensive Enablement

Risks related to intentional misuse of AI systems by bad actors for harmful purposes including disinformation, cyberattacks, and fraud.

Note: This table is meant to provide illustrative examples and is not intended to cover all types of risks nor every event included under each type.

Type of Risk	Incidents(s) / Hazard(s)	Incident Response Analysis and Recommendations
Scams & Fraud	• Anthropic reports misuse of AI in North Korean IT Schemes (August, 2025) • Microsoft reports disrupting $4 billion in AI-enabled fraud attempts in one year (April 2025)	• Currently, the FBI and FTC accept voluntary reporting from victims of scams and fraud. • **CAISI:** Establish incident reporting standards and data processing guidance that allows incident reports from both consumers, media, and industry to be aggregated in order to understand the scale and scope of harms and identify any common cause or vector.
Impersonation	• FBI warns of impersonation of senior officials, reported elsewhere to include the White House Chief of Staff and the Secretary of State (May 2025)	• Currently, the FBI and CISA have issued published guidance on avoiding impersonation scams and they accept voluntary reporting by victims. • **CAISI:** Adoption of secure and robust content authenticity standards would help authoritatively label official releases and content as legitimate.

Cyberattack	• LAMEHUG uses LLM to generate attack commands to run on victims' machines (July 2025)	• Currently, these AI incidents are relatively small-scale, experimental, or proofs-of-concepts. However, these are warning signs about weaknesses that could be exploited at a larger scale and become major security incidents.
		• Currently, cyber incident response resources include:
		  ◦ Voluntary Cyber Incident Reporting to CISA
		  ◦ Cybersecurity Incident Response Resources from CISA
		• **CAISI:** Establish common guidance for industry and academia, informed by general cybersecurity vulnerability disclosure, on AI risk disclosure best practices during testing, development, and deployment of AI systems.

Explosives	• Chatbot allegedly used to help craft explosives used in a deadly attack. (May 2025)	
Bioweapon Development	• Researchers find that AI-generated toxin sequences slip through DNA synthesis providers' screening tools (Oct 2025)	