CSET CENTER for SECURITY and EMERGING TECHNOLOGY

Key Takeaways From "The Mechanisms of AI Harm: Lessons Learned from AI Incidents"

Understanding how AI deployment can lead to harm is critical to improving risk mitigation. The number of recorded AI harm incidents has surged in recent years, suggesting that current approaches for AI harm prevention are falling short. This report argues that an insufficient understanding of how deployment can result in harm may be a barrier to comprehensive and effective mitigation strategies.

Drawing on data from the AI Incident Database, this report identifies and illustrates six different mechanisms of harm. Harms can result from intentional (e.g., harm by design, misuse, cyberattacks) and unintentional (e.g., AI failures, human-in-the-loop failures, integration harm) mechanisms, and involve a wide range of different stakeholders that must partake in mitigation efforts.

The analysis suggests that a one-size-fits-all approach to harm prevention will fall short. Effective risk mitigation requires diverse response strategies that address the various harm mechanisms. Model-based approaches alone, in particular, are insufficient to mitigate sociotechnical risks.

Risk mitigation strategies that rely on advanced model capabilities to proxy for risk are inadequate. Empirical evidence from incident data shows that to date, many instances of real-world harm result from specialized AI systems. While this trend may evolve as large, general-purpose AI models become more widely integrated in the economy and public administration, failure to target risks associated with the irresponsible design, development, and deployment of less powerful AI systems will result in continued harm.

Recommendations:

- AI developers and deployers should broaden their AI risk mitigation strategies to address all six AI harm mechanisms, especially when modeling how risks may materialize after deployment.
- Policymakers should establish an AI incident reporting regime to improve the government's capacity to respond to emerging risks and harm mechanisms and help overcome data limitations of current tracking efforts.

For more information:

- Download the report: https://cset.georgetown.edu/publication/the-mechanisms-of-ai-harm-lessons-learned-from-ai-incidents/
- Contact us: cset@georgetown.edu