Proposing A Bilateral U.S.-China Frontier AI Incident Reporting System

SUMMARY

There is an urgent need for more collaboration and dialogue between the U.S. and China on the security, cyber, and biological risks that artificial intelligence systems pose. AI incident reporting is a highly effective practice to document adverse effects during the deployment of AI systems. The proposed AI incident reporting system creates a feedback loop where stakeholders can identify patterns, implement corrections and prevent further failures.

PROBLEM

Neither the U.S. nor China, countries with advanced AI capabilities and development, have established a national or international AI-specific system for AI incident reporting due to mistrust and lack of dialogue. This creates the potential for small-scale AI failures to be misinterpreted as deliberate and adversarial actions, and for uncontrolled AI risks to escalate.

SOLUTION FRAMEWORK

- The main aim of the incident reporting system is to prevent escalation and further damage caused from the deployment of AI systems through information sharing.
- **Triggers**: Governments should report any serious incidents that cause human injury, monetary damages, or the disruption caused by an AI system that leads to (a) the death or serious harm to more than 1 person, (b) economic losses of more than $200k USD, or (c) a serious breach or disruption of critical infrastructure for over an hour.
- **Mechanism for incident reporting**: Reporting should happen within a 48-hour time window of the incident and occur through existing secure national points of contact.
- **Information sharing**: A short incident summary should be shared, including a model card, the nature of failure, containment status, an impact summary, and follow-up measures.
- Developers and stakeholders should mitigate the effects of this incident, create incident response plans, and establish deployment corrections as allowable actions.

ANALYSIS

- Due to the mistrust between the U.S. and China, we expect that this incident reporting system will be the most feasible when reporting is **voluntary, confidential, and non-punitive**. This also ensures that actors are incentivised to report incidents for safety without fear of retribution or penalties.
- Reports should be reviewed by a **joint working group** composed of developers from both countries to ensure fairness and oversight.
- To promote information sharing, the best way to ensure coordination would be through national institutions, such as the **US AISI and CNAISDA**. In the future, this incident reporting system is best coordinated through an international institute for AI security and lays the groundwork for an international AI safety body.