Summary of Adding Structure to AI Harm

Harms from the use of artificial intelligence systems ('AI harms') are varied and widespread. Monitoring and examining these harms is critical for mitigating risks from AI. By improving our understanding of how AI systems cause harm, enabling earlier detection of emerging types of harm, and directing resources to where prevention is needed most, AI harm analyses directly inform AI risk mitigation efforts.

The CSET AI Harm Framework identifies the key elements of AI harm, their relational structure and definitions without imposing a single interpretation of AI harm. Specifically, it:

- **Defines 'AI' harm** as when an entity experiences harm (or potential for harm) that is directly linked to the behavior of an AI system.
- **Groups harm from AI into either tangible or intangible harm.** Tangible harm is harm that is observable, verifiable, and definitive. Intangible harm is harm that cannot be directly observed or does not have any material or physical effect.
- **Allows users to define additional categories of tangible and intangible harm.** The framework provides some common categories of harm and also allows the inclusion of new categories.
- **Distinguishes harm that actually occurred from harm that may occur.** This allows for the tracking of definitively experienced harm while also enabling research and analysis on harm risk and vulnerability.

This paper provides additional information about the framework. It:

- **Discusses how users can adapt the framework.** Users can create a customized taxonomy, modifying the base modular structure and providing definitions.
- **Provides an example customized framework.** We illustrate the customization process using CSET's AI Harm Taxonomy for the AI Incident Database.
- **Details future additions to the framework.**

For more information:

- Download the report: https://cset.georgetown.edu/publications/adding-structure-to-ai-harm
- Contact Us: Mia Hoffmann, mia.hoffmann@georgetown.edu