


IST's Efforts in an Age of AI: An Overview

We work with a diverse range of stakeholders across the AI ecosystem to produce:


risk mitigation strategies

Ongoing Monitoring and Maintenance	Types of risks mitigated
Reporting mechanisms Establish clear and widely accessible reporting mechanisms for individuals to report suspected fraud schemes or malicious use of AI technologies. → Accessible reporting mechanisms help quickly identify and address misuse.	
Public campaigns Ensure that reporting mechanisms are well-publicized and easily discoverable through various media outlets, public awareness campaigns, and educational materials. <small>→ Building trust in reporting mechanisms is a key goal of our work.</small>	


useful tools and frameworks



recommendations



CBMs that involve agreeing to, or communicating an intent to, renounce or limit the use of AI technologies in certain weapon and military systems.



CBMs that encourage governments and industry players to agree on standards, guidelines, and norms related to AI trust and safety, as well as "responsible" use of AI technologies.

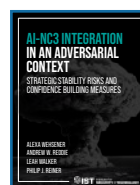
Implications of AI for national security and global stability: IST's efforts to understand the implications of AI began in 2017, with a roundtable featuring leading developers and a workshop on AI's societal implications. In 2018, IST undertook a joint initiative with Lawrence Livermore National Laboratory's Center for Global Security Research that aimed to articulate, understand and manage the long-term opportunities and risks posed by AI-related technologies for international security, global stability, and warfare.



AI risk reduction amid the implications of openness: With the support of the Patrick J. McGovern Foundation, IST is engaging with stakeholders to craft useful tools and frameworks for understanding how access to AI foundation models and their components impacts the risk they pose to individuals, groups, and society. In 2023, IST designed a novel matrix to map categories of risk against a gradient of access to AI foundation models. A subsequent report established a lifecycle approach to AI risk reduction, along with 5 guiding principles for risk mitigation, and applied the framework to the risk of malicious use to determine effective risk mitigation strategies.



AI and ML integration into nuclear command, control and communications (NC3) systems: In January 2019, in collaboration with the Nautilus Institute for Security and Sustainability and Stanford University's Preventive Defense Project, IST hosted a multi-stakeholder discussion on the modernization of global NC3 systems. Building on this foundation, IST convened scientists, engineers, policymakers, and academics to examine policy tools that could mitigate the risks posed by the integration of AI into NC3 systems. With the support of the State Department's Bureau of Arms Control, Verification, and Compliance, IST proposed confidence-building measures to limit the use of AI in weapon systems, encourage the creation of norms around the use of AI, increase lines of communication, and bolster collaboration between private industry and government.



Implications of AI in cybersecurity: With the support of Google.org through its Digital Futures Project, IST is studying the applications of AI in cybersecurity and implications for the offense-defense balance. IST aims to provide a clear picture of current cybersecurity trends, cutting through marketing hype to offer a future outlook and actionable recommendations. This effort is part of a broader IST project to identify key cybersecurity areas needing focus, such as threat intelligence, automated defenses, and scalable security solutions. IST also co-leads a complementary effort with the World Economic Forum's Centre for Cybersecurity to understand the implications of AI in the cybercrime ecosystem.

**Get in touch
with us**

WEBSITE
→ www.securityandtechnology.org

LINKEDIN
@Institute for Security and Technology