



ADVANCED CYBER SECURITY CENTER

Trusted Networks. Advancing Cyber Strategies.

2nd Annual

Cyber Risk Governance

WORKSHOP

Thank you

Workshop Host



Event Sponsor



CRG Lead Research Partner



Emerging Tech Research Partners



AI Risks, Threats and Opportunities



Mark Maybury
Lockheed Martin



Marc Zissman
MIT Lincoln Laboratory

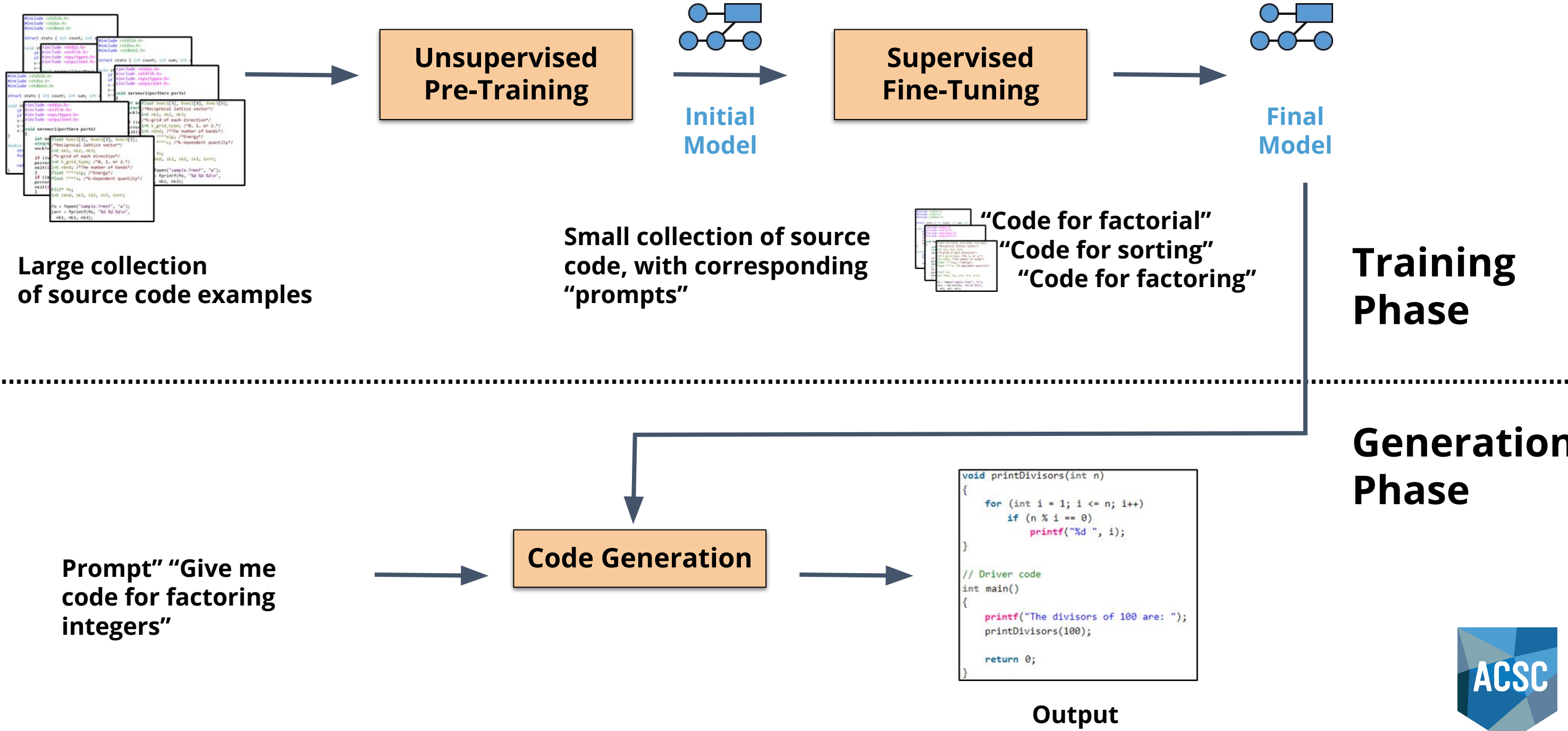
Promise of AI and GAI



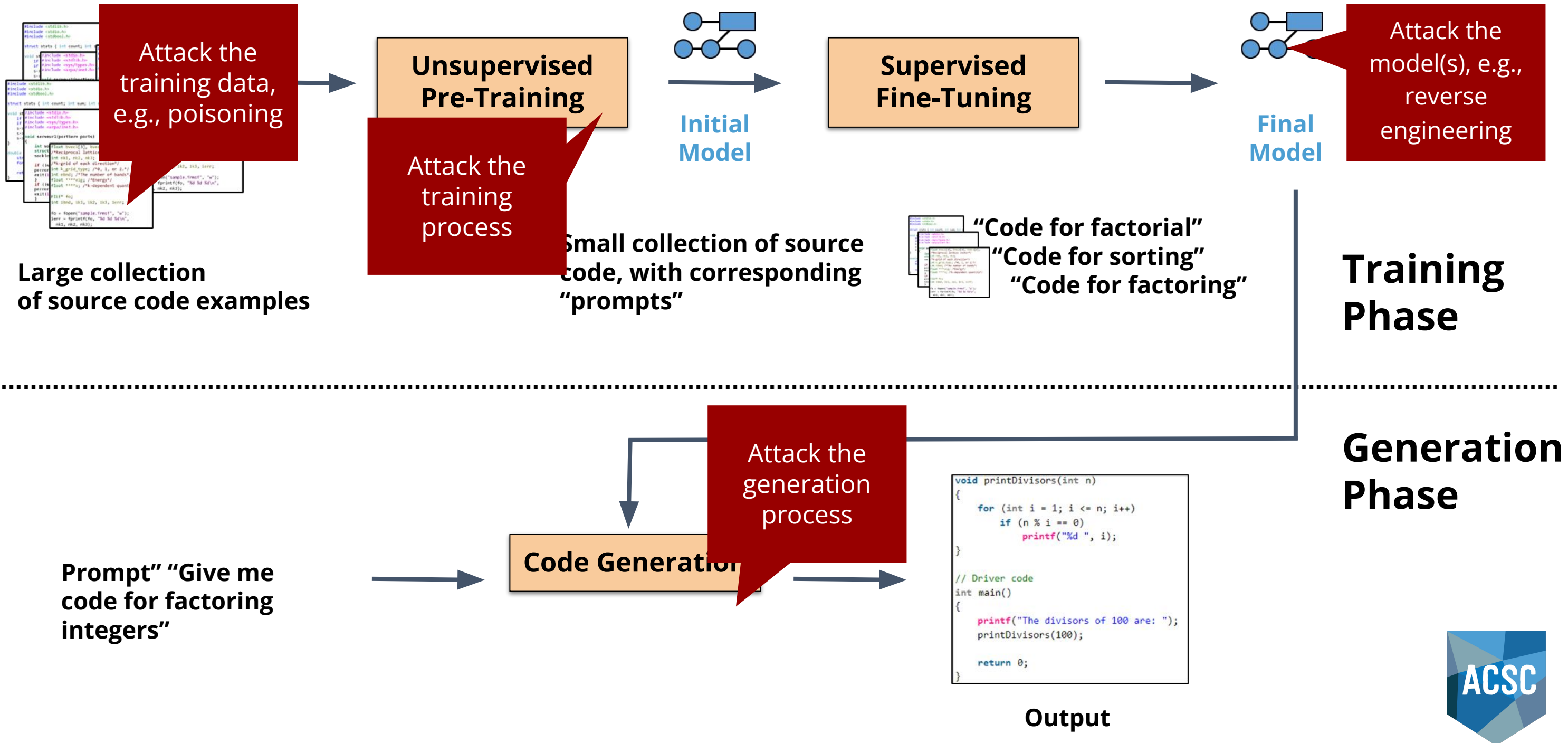
DARPA Semantic Forensics Video



A GenAI Application: Software Development



Attacking the Software Development System



GenAI Threats - Examples

		Threats from an adversary attacking my use of GenAI	Threats from an adversary using GenAI against me
Conventional Cyber Threats	Confidentiality	<ul style="list-style-type: none"> • Training data are exfiltrated • Models are exfiltrated • Models are reverse engineered, exposing IP or PHI or PII • IP or PHI or PII leaked into commercial GenAI system 	<ul style="list-style-type: none"> • Social engineering at speed, scale and with high-fidelity to gain unauthorized access
	Integrity	<ul style="list-style-type: none"> • Training data are poisoned • Models are corrupted 	<ul style="list-style-type: none"> • Fake training data generated with high-fidelity at speed and scale leading to poor performance of other GenAI systems
	Availability	<ul style="list-style-type: none"> • Data and/or models are held for ransom 	<ul style="list-style-type: none"> • Extremely high-fidelity email and web site flooding at speed and scale
GenAI-specific Threats		<ul style="list-style-type: none"> • System underperforms due to inadequate training or incomplete testing • IP infringement due to insufficient licensing of training data • Lack of IP protecting for GenAI-generated material • GenAI output misleads recipients into thinking it was human-generated 	<ul style="list-style-type: none"> • Disinformation campaign at hitherto unexperienced speed and scale

Successful attacks could have reputational, financial and legal consequences

DeepFakes Arms Race

- **BBC [Deepfake Quiz](#)** – increasingly hard for humans to detect fakes
- **Recent Examples:**
 - [Fake Pentagon Explosion](#) via verified Twitter account, retweeted by RT and others, market drops \$500B
 - [Fake Biden Robo Call](#) in NH Primary
- **Deepfakes Study (Nature 2024)**
 - **Test subjects:** Nationally representative panels from RAND Corp, 761 unique 18+ year old individuals from the American Life Panel (ALP), 740 from K-12 principals at American School Leader Panel (ASLP) ASLP, and 642 teachers from the American Teacher Panel (ATP).
 - **Results: 27-50% of people cannot distinguish authentic from deepfake videos**
 - **Vulnerability increases with age and trust of info sources**
 - **Adults and educators most susceptible compared to students**
 - **Vulnerability increases with exposure!**
- **Reality Defender (Ben Colman) detects deepfakes (they contrast nation state capability vs. criminal “cheapfakes”)** for clients like Taiwanese government, NATO, media organizations and large banks.

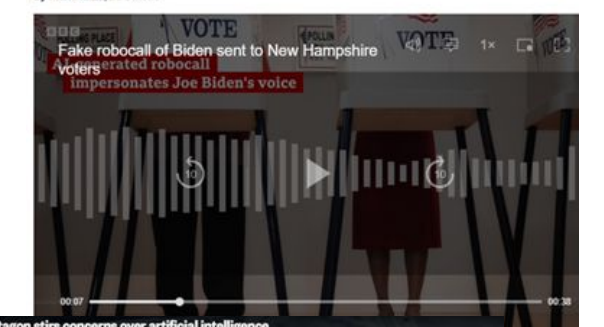
BBC News Source: <https://apple.news/ACZ04WepgSPKqiPmsvFiHUQ>



Fake Biden robocall tells voters to skip New Hampshire primary election

22 January 2024

By Max Matza, BBC News



Transparency Example: GAI Model Cards

- **Why:** Promote transparency and accountability (ref “Nutritional Labels”)
- **What**
 - Model details: version, training and test data, licenses, contacts
 - Intended use
 - Performance metrics
 - Evaluation procedures e.g., Voter demographics: performance across culture, race, age, gender, location
 - Ethics: Privacy, fairness, potential harms
- **Who:** Policymakers, organizations, developers, privacy/security experts, end users
- **Example Model Cards**
 - Meta and Microsoft: [Llama 2, research paper](#)
 - OpenAI’s [GPT-3 model card, research paper](#)
 - Google’s [face detection model card,](#)
 - IBM [AI Factsheet](#)



References:

Desai, A. [5 things to know about AI model cards](#)

[A “Nutrition label” for privacy, 5th Symposium on Usable Privacy and Security, 2009.](#)



Some Threat Mitigation Strategies

- Develop **appropriate policies** for
 - Use of commercial, cloud-based GenAI systems
 - Assessment of the claims made by the GenAI system vendors
 - Marking, tracking, testing of IP produced by GenAI systems
- Acquire and use **local GenAI systems for sensitive applications**
- Track evolving legal landscape re: **IP infringement of web-scraped data**
- Train **staff to be vigilant re: deep fakes**
- Maintain **heightened vigilance re: cyber attacks on GenAI systems**



Summary

- **AI-enabled systems are taking the world by storm**
- **Risks are manifest**
 - New threats on C, I and A
 - New threats that are specific to GenAI
- **Risk frameworks are nascent and probably incomplete**
 - E.g., MITRE ATLAS (atlas.mitre.org) and NIST AI RMF*
 - Not obvious (yet) how to prioritize limited mitigation resources against various types of risk
- **All levels of the organization**, from practitioners up to the board, need to work to **mitigate the risks of AI and GenAI while exploiting the value**

* See <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

