2024 CRA CONFERENCE

The Security Risks of Generative AI: From Identification and Mitigation to Responsible Use

Mihai Christodorescu
Google

Somesh Jha
University of Wisconsin

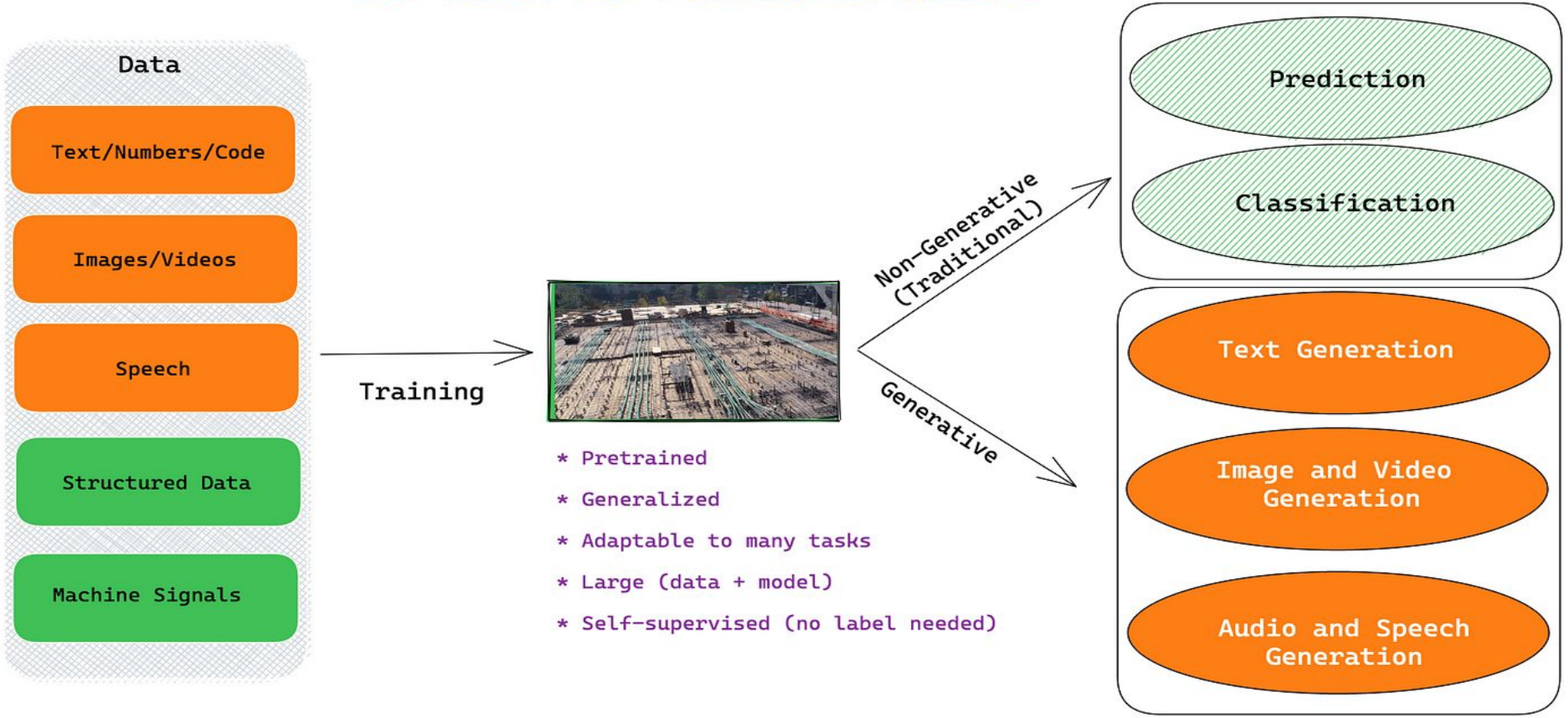
John Mitchell
Stanford University

Rebecca Wright
Barnard College

Matt Turek
DARPA

GenAI Performant in Multiple Contexts

Generative and Non-generative Use Cases for Foundation Models



Dual Use
according to Wikipedia

2024 CRA CONFERENCE

- … **dual-use items** refers to goods, software and technology that can be used for both civilian and military applications
- The “dual-use dilemma” was first noted with the discovery of the process for synthesizing and mass-producing ammonia which revolutionized agriculture with modern fertilizers but also led to the creation of chemical weapons during World War I.

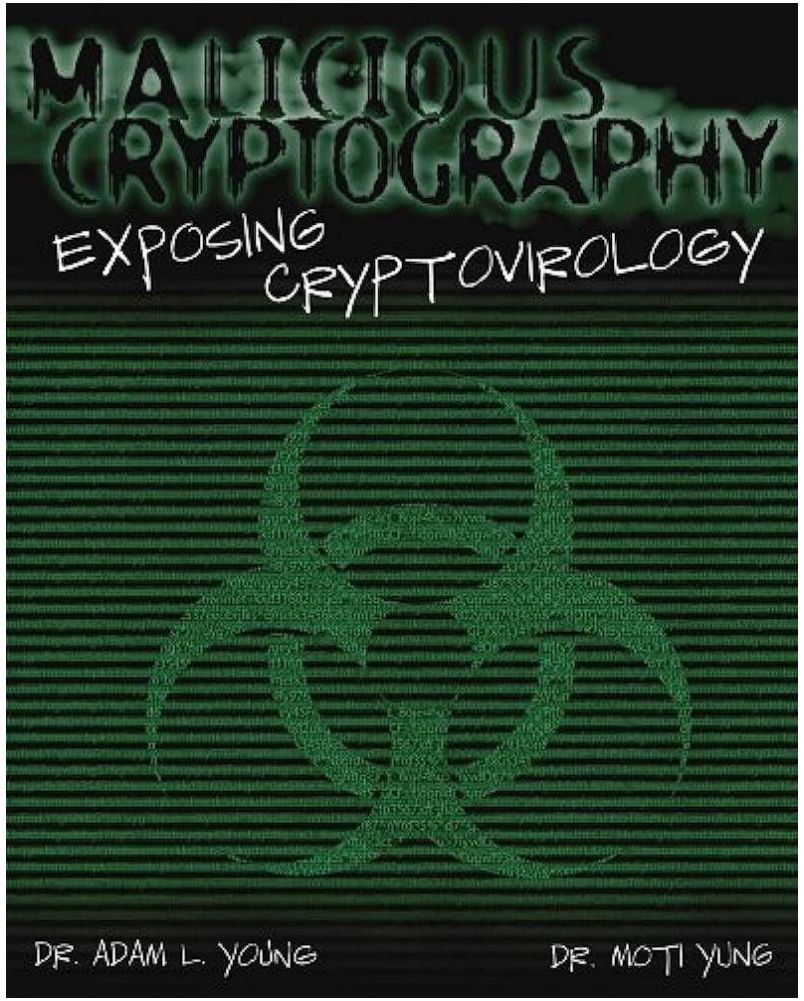
Dual Use in Cryptography

2024 CRA CONFERENCE

- Cryptography has very important use cases
  - Secret communication
  - Encrypting data
  - ….

- Adversaries can also use Cryptography
  - Secret communication among bad actors
  - Ransomware: encrypting files
  - …

Want to be scared?



GenAI: Transformational? Risky?

2024 CRA CONFERENCE

GenAI amplifies creativity and productivity

Example: all aspects of programming enhanced by GenAI

- Code generation, code understanding, code testing, code repair, ...
- Huge change to how software developers work

GenAI is not yet trustworthy

- Bad actors also benefit from the power of GenAI
- Open problem: making GenAI and the systems around GenAI safe and secure

Previous Events

2024 CRA CONFERENCE

June 2023

Oct 2023

May 2024

Papers based on workshops:

Identifying and Mitigating the Security Risks of Generative AI, Foundations and Trends in Privacy and Security, Vol 6, No1, Dec 2023

Identifying and Mitigating the Security Risks of Generative AI
https://arxiv.org/abs/2308.14840

Today's Panel

Which GenAI directions are most exciting? How to realize them?

Which GenAI risks are most likely? How to mitigate them?

What is the role for the computing research community?

Intro: Rebecca Wright

2024 CRA CONFERENCE

Druckenmiller Professor and Chair of Computer Science, Barnard College
Director, Vagelos Computational Science Center
Chair, Cybersecurity Research Center, Data Science Institute, Columbia University
previously **Professor of Computer Science**, **Director of DIMACS** at Rutgers

Research Interests

- Computer and communications security
- Privacy
- Cryptographic protocols
- Fault-tolerant distributed computing

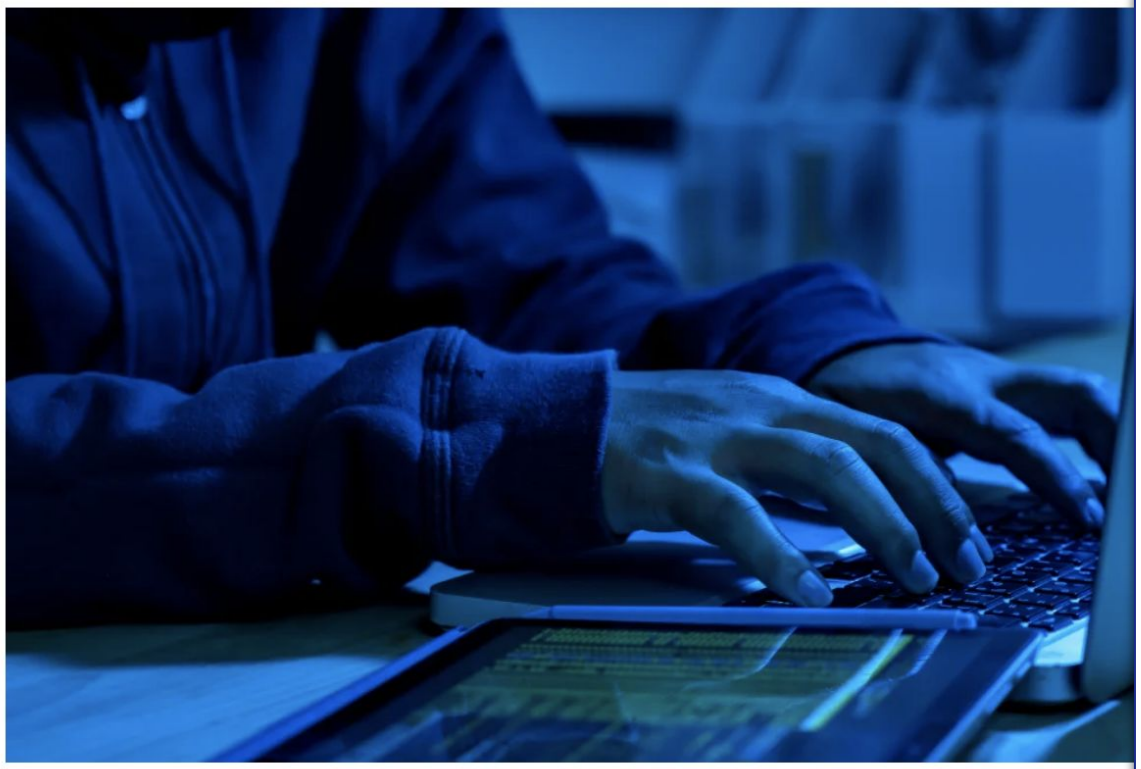
Intro: Rebecca Wright (Barnard College)

- Like any new widely adopted technology, GenAI brings new threats that we haven't yet addressed. Risks include applying old mindsets and old modes of thinking without understanding new contexts.
- GenAI enhances attackers' ability to carry out attacks, including sociotechnical attacks that try to get people to do things, as well as generating malicious code.
- Risks to privacy because so much data is needed to train models. Also, models can reveal their sensitive information from their training data.
- New basic research and translational/practical research are needed. Ex: if we had perfect detectors for AI-generated content, how would we use them to protect cybersecurity and protect people while still allowing desired uses?

Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’

By Heather Chen and Kathleen Magramo, CNN

2 minute read · Published 2:31 AM EST, Sun February 4, 2024



Authorities are increasingly concerned at the damaging potential posed by artificial intelligence technology. boonchai wedmakawand/Moment RF/Getty Images

(CNN) — A finance worker at a multinational firm was tricked into paying out $25 million to fraudsters using deepfake technology to pose as the company’s chief financial officer in a

TECHNOLOGY

That Colleague or Customer on Zoom Might Be an AI Deepfake. Here's How You Can Tell

Think it can't happen? A Hong Kong company just lost $25.6 million to a deepfake version of its CFO.

EXPERT OPINION BY MINDA ZETLIN, AUTHOR OF 'CAREER SELF-CARE: FIND YOUR HAPPINESS, SUCCESS, AND FULFILLMENT AT WORK' @MINDAZETLIN

FEB 8, 2024

Intro: John Mitchell

Mary and Gordon Crary Family Professor of Computer Science and (by courtesy) Electrical Engineering and Education, Stanford University

previously **Stanford Vice Provost for Teaching and Learning** and **chair of the Computer Science Department**

Research Interests

- Programming languages
- Computer security and privacy
- Blockchain
- Trustworthy machine learning
- Technology for education

Intro: John Mitchell (Stanford University)

2024 CRA CONFERENCE

- AI is hugely effective for many tasks
  - Programming has been transformed
  - Creativity is enhanced: write spearfishing email, create deep fakes
  - Productivity is expanded: data analysis, summarization, workflow mgmt,...
- AI is not that trustworthy, a decades-long challenge
- It's human nature to do both good and bad
  - Bad actors now more powerful
  - We face a challenge to develop new defenses
- Many examples
  - Education: provide useful encouragement without harm
  - Web security: protect applications that rely on AI

Intro: Matt Turek

Deputy Director, Information Innovation Office (I2O), DARPA

previously **Program Manager** for Media Forensics (MediFor), Semantic Forensics (SemaFor), Machine Common Sense (MCS), Explainable AI (XAI), and Reverse Engineering of Deception (RED) AI Exploration program (AIE) programs

Research Interests

- Computer vision
- Machine learning
- Artificial intelligence
- CV/ML/AI applications to problems with significant societal impact

Today's Panel

Which GenAI directions are most exciting? How to realize them?

Which GenAI risks are most likely? How to mitigate them?

What is the role for the computing research community?

2024 CRA CONFERENCE

The Security Risks of Generative AI: From Identification and Mitigation to Responsible Use

Mihai Christodorescu
Google

Somesh Jha
University of Wisconsin

John Mitchell
Stanford University

Rebecca Wright
Barnard College

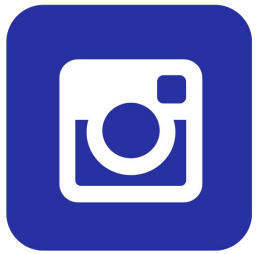
Matt Turek
DARPA

2024 CRA CONFERENCE

Stay connected!

Follow CRA on social media

@computing-research-association



@computingresearch

@computingresearch

Attacks Leveraging GenAI

2024 CRA CONFERENCE

- Spear-phishing
- Deepfakes
- Proliferation of cyberattacks
- Low barrier-to-entry for adversaries
  - WormGPT, FraudGPT…
- *Hallucinations (*)*
- *Lack of social awareness and human sensibility (*)*
- *Data feedback loops (*)*
- *Unpredictability (*)*

(*): Inherent limitations that can be exploited by the attacker

Fake News in Elections!

2024 CRA CONFERENCE

Really Interesting/Scary Paper..

- Nicholas Carlini (Google DeepMind)
  "A LLM Assisted Exploitation of AI-Guardian"

As a case study, we evaluate the robustness of AI-Guardian, a recent defense to adversarial examples published at IEEE S&P 2023, a top computer security conference

....

We write none of the code to attack this model, and instead prompt GPT-4 to implement all attack algorithms following our instructions and guidance.