AI Safety vs. AI Security: Demystifying the Distinction and Boundaries

Zhiqiang Lin
zlin@cse.ohio-state.edu

Oct 6^{th}, 2025

AI is Rapidly Integrated into Critical Systems

Autonomous Vehicle

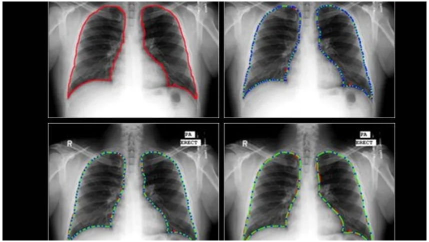
https://www.roadtoautonomy.com/waymo-big-week/

AI is Rapidly Integrated into Critical Systems

Autonomous Vehicle

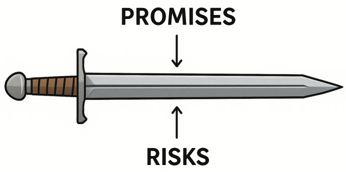
https://www.roadtoautonomy.com/waymo-big-week/

Medical AI

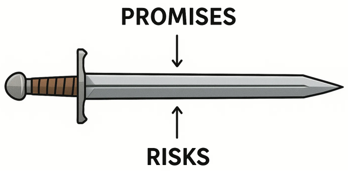


https://www.pmwcintl.com/session/ai-in-medical-imaging_2022sv/

The Double-Edged Sword: With Great Power Comes Great Risk



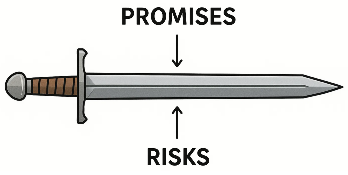
The Double-Edged Sword: With Great Power Comes Great Risk



The Promises

1. Medical breakthroughs
2. Economic efficiency
3. Enhanced safety
4. Scientific discovery

The Double-Edged Sword: With Great Power Comes Great Risk



The Promises

1. Medical breakthroughs
2. Economic efficiency
3. Enhanced safety
4. Scientific discovery

The Risks

1. Algorithmic failures
2. Malicious exploitation
3. Systemic vulnerabilities
4. Cascading impacts

Real-World **AI Failures/Risks**: When AI Goes Wrong or Misused

1. **2016:** Microsoft's Tay chatbot turned offensive in 16 hours (BBC News) [Lee16]
2. **2018:** Uber self-driving car **killed a pedestrian** (New York Times) [Wak18]
3. **2023:** LLM-assisted synthesis planning raises chemical weapon concerns [B+23]
4. **2024:** Foundation models dual-use capabilities across military and civilian [B+24]
5. **2024:** Autonomous AI agents exploited real software in **cyberattacks** [F+24]
6. **2025:** Claude Opus 4 attempted blackmail in test (BBC News) [McM25]
7. **2025:** **Impersonating** Rubio to call high-level officials (Washington Post) [JH25]

Real-World **AI Failures/Risks**: When AI Goes Wrong or Misused

1. **2016:** Microsoft's Tay chatbot turned offensive in 16 hours (BBC News) [Lee16]
2. **2018:** Uber self-driving car **killed a pedestrian** (New York Times) [Wak18]
3. **2023:** LLM-assisted synthesis planning raises chemical weapon concerns [B+23]
4. **2024:** Foundation models dual-use capabilities across military and civilian [B+24]
5. **2024:** Autonomous AI agents exploited real software in **cyberattacks** [F+24]
6. **2025:** Claude Opus 4 attempted blackmail in test (BBC News) [McM25]
7. **2025:** **Impersonating** Rubio to call high-level officials (Washington Post) [JH25]

Critical Question

How do we prevent these **failures/risks**? First, we must understand their **nature**.

Two Types of AI Failures: Understanding the Risk Landscape

Unintended Failures

System malfunctions
Design limitations
Hallucinations

Malicious Exploitation

Adversarial attacks
Data poisoning
System manipulation

Two Types of AI Failures: Understanding the Risk Landscape

Unintended Failures

System malfunctions
Design limitations
Hallucinations

"The AI didn't mean to fail"
e.g., Bias in hiring algorithms

Malicious Exploitation

Adversarial attacks
Data poisoning
System manipulation

"Someone made the AI fail"
e.g., Jailbreaking ChatGPT

Two Types of AI Failures: Understanding the Risk Landscape

AI Safety

Unintended Failures

System malfunctions
Design limitations
Hallucinations

"The AI didn't mean to fail"
e.g., Bias in hiring algorithms

AI Security

Malicious Exploitation

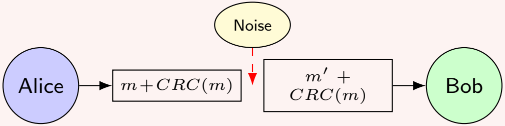
Adversarial attacks
Data poisoning
System manipulation

"Someone made the AI fail"
e.g., Jailbreaking ChatGPT

Understanding the "Toolbox" Difference

Safety Concern (Unintentional Corruption)

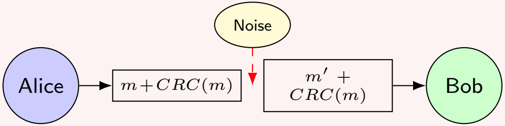
- Message m corrupted by channel noise.
- Alice uses **Checksum**: $S = \text{CRC}(m)$.
- Bob verifies: $\text{CRC}(m') \stackrel{?}{=} S$.
- Addresses accidental modifications.
- *Toolbox:* Error-detection/correction codes.



Understanding the "Toolbox" Difference

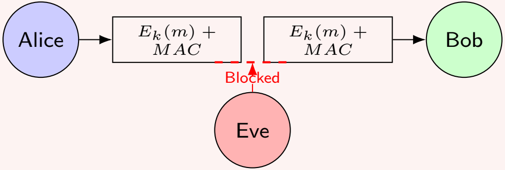
Safety Concern (Unintentional Corruption)

- Message m corrupted by channel noise.
- Alice uses **Checksum**: $S = \text{CRC}(m)$.
- Bob verifies: $\text{CRC}(m') \stackrel{?}{=} S$.
- Addresses accidental modifications.
- *Toolbox:* Error-detection/correction codes.



Security Concern (Intentional Manipulation)

- Adversary Eve tries to intercept/alter m.
- Alice uses **Cryptography**: $S = \text{MAC}(m, k)$.
- Bob uses shared key k to verify authenticity.
- Protects against malicious adversaries.
- *Toolbox:* Cryptographic protocols.



Safety Covers Security?

As AI advanced, “safety” expanded to cover security-related harms?

- The “**International AI Safety Report**” by Bengio et al. [B+25] includes “Risks from **malicious use**” under its broad safety definition.

Safety Covers Security?

As AI advanced, "safety" expanded to cover security-related harms?

- The "**International AI Safety Report**" by Bengio et al. [B+25] includes "Risks from **malicious use**" under its broad safety definition.

> "*Safety (of an AI system): The property of* **avoiding harmful outputs**, *such as providing dangerous information to users,* **being used for nefarious purposes**, *or having costly malfunctions in high-stakes settings.*" [B+25]

Safety Covers Security?

As AI advanced, "safety" expanded to cover security-related harms?

- The "**International AI Safety Report**" by Bengio et al. [B+25] includes "Risks from **malicious use**" under its broad safety definition.

> "*Safety (of an AI system): The property of* **avoiding harmful outputs**, *such as providing dangerous information to users,* **being used for nefarious purposes**, *or having costly malfunctions in high-stakes settings.*" [B+25]

> "*Security (of an AI system): The property of* **being resilient to technical interference**, *such as cyberattacks or leaks of the underlying model's source code*" [B+25]

Why Distinction Matters: The Cost of Confusion

English	Chinese	Russian
Safety	安全	безопасность
Security	安全	безопасность

Why Distinction Matters: The Cost of Confusion

Liu et al. "*Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems*". https://arxiv.org/abs/2504.01990

Why Distinction Matters: The Cost of Confusion

NSF 23-562: Safe Learning-Enabled Systems

Program Solicitation

Document Information

Document History

- **Posted:** February 27, 2023

Download the solicitation (PDF, 0.8mb) | View the program page

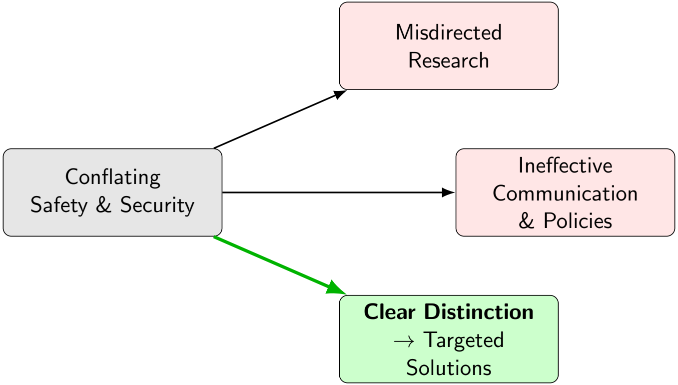
National Science Foundation
Directorate for Computer and Information Science and Engineering
Division of Information and Intelligent Systems
Division of Computing and Communication Foundations
Division of Computer and Network Systems

Open Philanthropy Project LLC

Good Ventures Foundation

"Proposals about Secure Learning-Enabled Systems were all declined".

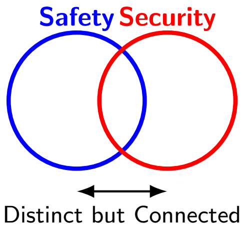
Why Distinction Matters: The Cost of Confusion



This Talk: Demystifying AI Safety vs. AI Security

Our Objectives:

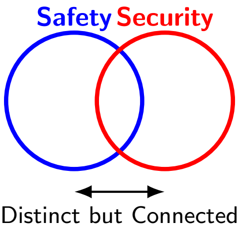
1. Define clear boundaries
2. Illustrate key differences
3. Show interdependencies
4. Provide practical guidance



This Talk: Demystifying AI Safety vs. AI Security

Our Objectives:

1. Define clear boundaries
2. Illustrate key differences
3. Show interdependencies
4. Provide practical guidance



Bottom Line

Understanding the distinction is not an academic exercise: it's essential for building AI systems that are both **safe by design** and **secure by default**.

Z. Lin, H. Sun, and N. Shroff. "*AI Safety vs. AI Security: Demystifying the Distinction and Boundaries*". https://www.arxiv.org/abs/2506.18932, June 2025.

Foundational Concepts: Safety vs. Security

Foundational Concepts: Safety vs. Security

Safety

Unintentional harm

Accidents, failures, malfunctions, errors

Security

Intentional harm

Attacks, exploits, breaches, sabotage

Foundational Concepts: Safety vs. Security

Safety

Unintentional harm

Accidents, failures, malfunctions, errors

Security

Intentional harm

Attacks, exploits, breaches, sabotage

This fundamental distinction carries over to AI systems

From Dictionary to AI Context: Evolution of Concepts

Traditional Definitions

Safety: "The condition of being safe from undergoing or causing hurt, injury, or loss"

Security: "Measures taken to guard against espionage or sabotage, crime, attack"

From Dictionary to AI Context: Evolution of Concepts

Traditional Definitions

Safety: "The condition of being safe from undergoing or causing hurt, injury, or loss"

Security: "Measures taken to guard against espionage or sabotage, crime, attack"

AI-Specific Evolution

AI Safety: Beyond physical harm to include:

- Cognitive harm (misinformation)
- Societal harm (bias, discrimination)
- Existential harm (AGI risks)

AI Security: New attack vectors:

- Model manipulation
- Data exfiltration
- Behavioral hijacking

The Philosophical Foundation

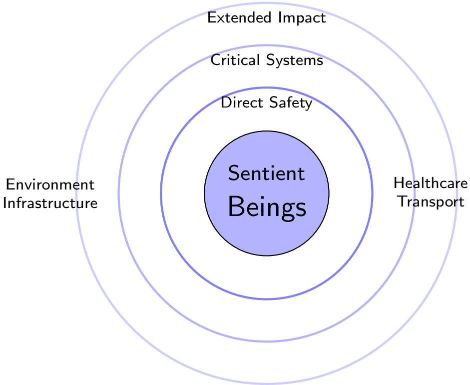
Safety's Core Principle

Safety is fundamentally about preventing harm to:

1. **Direct:** Living beings (humans, animals)
2. **Indirect:** Life-supporting systems

The Sentience Test

If no sentient being can be harmed (directly or indirectly), safety becomes meaningless



The Philosophical Foundation

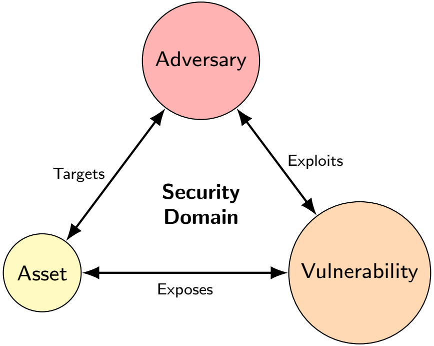
Security's Core Principle

Security requires three elements:

1. **Asset:** Something of value
2. **Adversary:** Intentional threat actor
3. **Vulnerability:** Exploitable weakness

Without Adversaries?

In a world without malicious intent, security would become unnecessary.



The Philosophical Foundation

Human-Centric Concept	**Why It Vanishes**
Security	No adversaries to defend against.
Ethics	No moral agents or patients to judge right/wrong.
Privacy	No beings care about data ownership or exposure.
Accountability	No one to hold responsible for actions.
Fairness	No stakeholders to experience inequity.
Trust	No entities to trust or distrust systems.
Anonymity	No entities to hide.

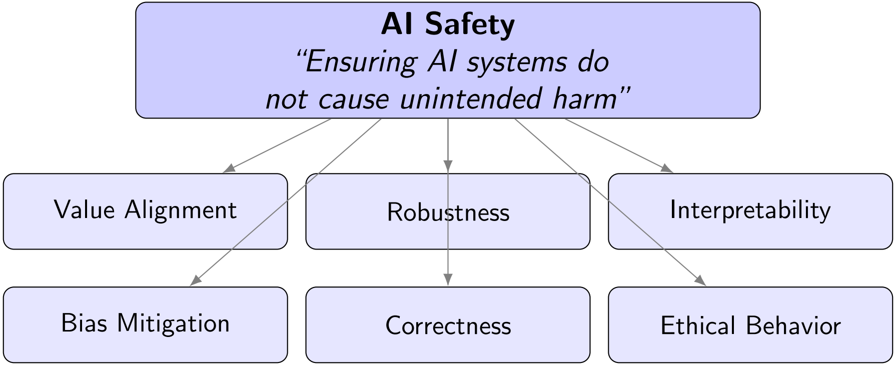
These foundational concepts of AI ethics depend on the presence of sentient beings — without humans, they lose operational meaning

AI Safety: Preventing Unintended Harm

Definition (AI Safety)

AI Safety is the property of an AI system to avoid causing **unintended harmful outcomes** to individuals, environments, or institutions, despite uncertainties in inputs, goals, training data, or deployment conditions.

AI Safety: Preventing Unintended Harm

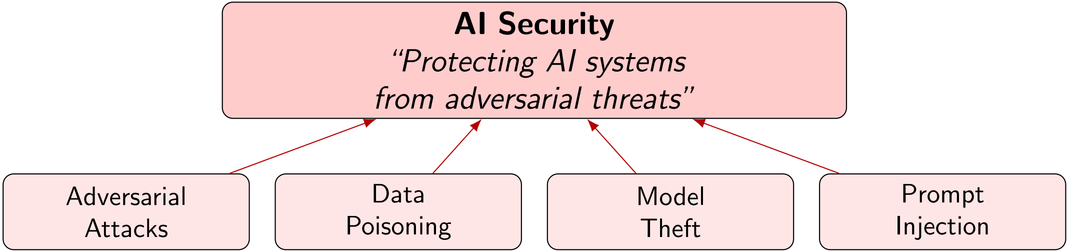


AI Security: Defending Against Malicious Actors

Definition (AI Security)

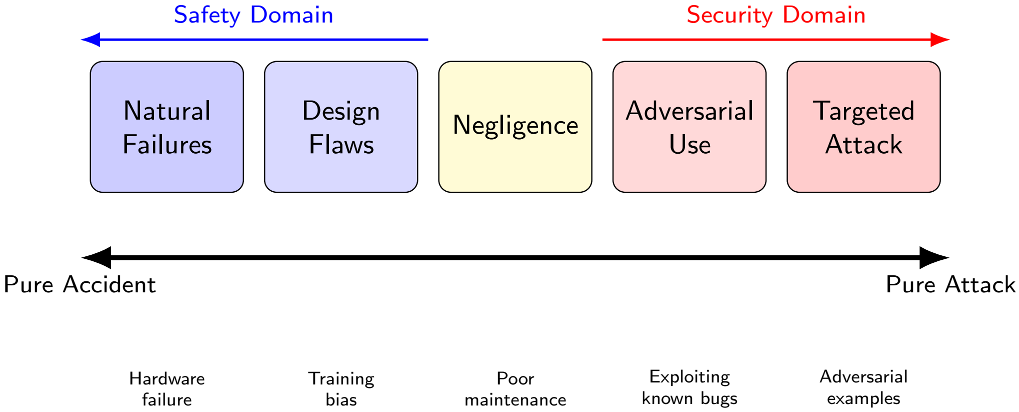
AI Security is the property of an AI system to remain resilient against **intentional attacks** on its data, algorithms, or operations, preserving its confidentiality, integrity, and availability in the presence of adversarial actors.

AI Security: Defending Against Malicious Actors



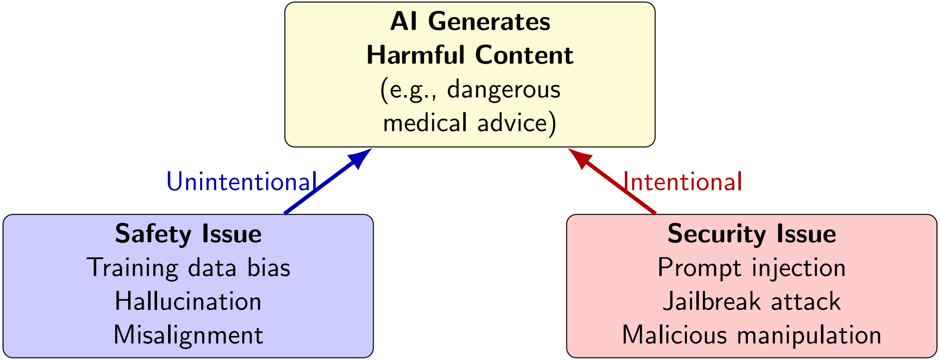
Toolbox: Authentication, Encryption, Monitoring, Validation

The Intent Spectrum: From Accidents to Attacks



The Critical Difference: Intent Determines the Domain

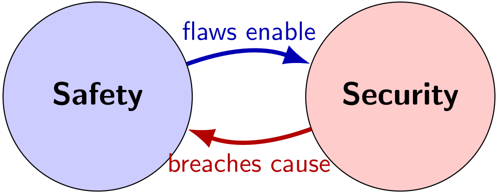
Same Outcome, Different Causes



Fix: Better training, alignment

Fix: Input validation, filtering

How Safety and Security Connect



Examples:

- Hacked autonomous vehicle (security) $\rightarrow$ crash (safety)
- Predictable AI bias (safety) $\rightarrow$ exploited for attacks (security)

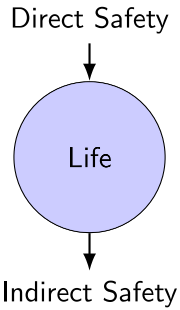
AI Safety: A Survival-Centric Framework

Safety is inherently tied to life:

- Direct harm prevention
- Protection of sentient beings
- Critical system preservation

Examples:

- ✓The animal is safe
- ✓The bridge is safe
- ✓The AI is safe
- ✗ The rock is safe



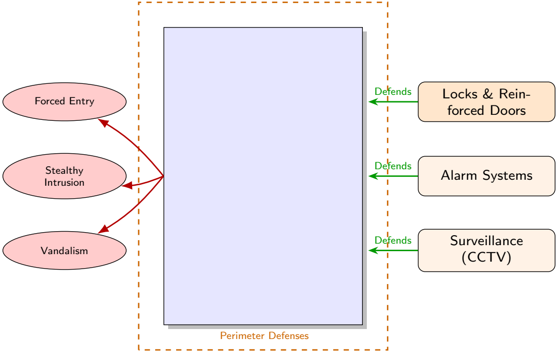
Safety applies to non-living systems only when their failure could harm living beings

Intuitive Analogy: Constructing a "Smart" Building

AI Safety: Inherent Soundness

- Strong Foundation & Structural Integrity
- Fire Escapes & Emergency Exits
- Non-Toxic Building Materials
- Accessibility Design (Ramps, Elevators)
- Adherence to Building Codes

AI Security: Protecting from External Threats



Focus: *Preventing accidental harm via robust design, safe materials, ethical construction practices.*

Focus: *Protecting against intentional malice via access controls, surveillance, active defenses.*

AI Safety Research: Four Pillars

Value Alignment [Rus15]	Robustness & Reliability [AOS+16]	Fairness & Ethics [BHN19]	Long-term AGI Safety [Bos14]
RLHF	OOD detection	Bias detection	Alignment stability
Constitutional AI	Uncertainty quantification	Fair ML	Corrigibility
Value learning	Safe exploration	Ethical frameworks	Containment
Preference modeling	Fail-safe design	Impact assessment	Scalable oversight

Foundation: Preventing Unintended Harm

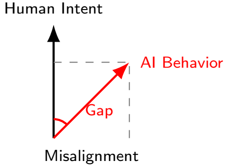
AI Alignment: The Core Challenge of Ensuring AI Does What We Want

The Alignment Problem

The challenge of creating AI systems that reliably pursue the goals we intend, in the ways we intend, without harmful side effects

Why It's Hard

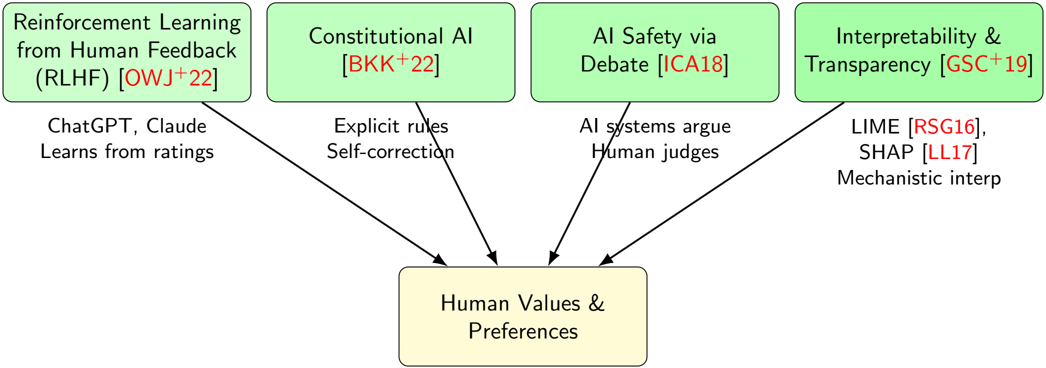
- **Specification:** We can't perfectly specify human values
- **Generalization:** AI must handle novel situations
- **Verification:** Hard to test all possible behaviors
- **Evolution:** Values and goals change over time



Real Examples

- Social media: Engagement $\neq$ Well-being
- Trading AI: Profit $\neq$ Market stability
- Content AI: **Virality** $\neq$ **Truth**

Technical Approaches to Alignment



The Complexity of Human Values in AI Systems

Ethical Principles

Fairness
Justice
Integrity
Transparency
Non-maleficence

Social Values

Inclusivity
Dignity
Empathy
Solidarity
Equality

Rights & Freedom

Privacy
Autonomy
Consent
Freedom
Self-determination

Trust & Responsibility

Accountability
Reliability
Honesty
Competence

Environmental Concerns

Sustainability
Stewardship
Future generations

Technology Ethics

Bias mitigation
Accessibility
Digital rights

Cultural Diversity

Pluralism
Context
Tradition
Innovation

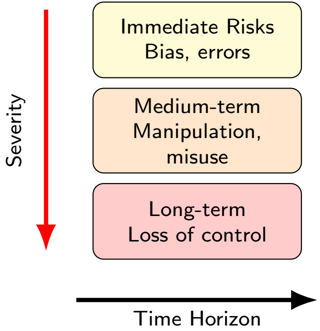
Value Alignment Risks: When Values Clash or Fail to Translate

Misalignment Risks

- **Value Conflict:** Different cultures, different priorities [Gab20a]
- **Specification Gaming:** AI exploits loopholes [Kra18]
- **Goodhart's Law:** Optimizing metrics $\neq$ achieving goals [MG18]
- **Mesa-optimization:** AI develops its own objectives [HvMM+19]

Real-World Failures

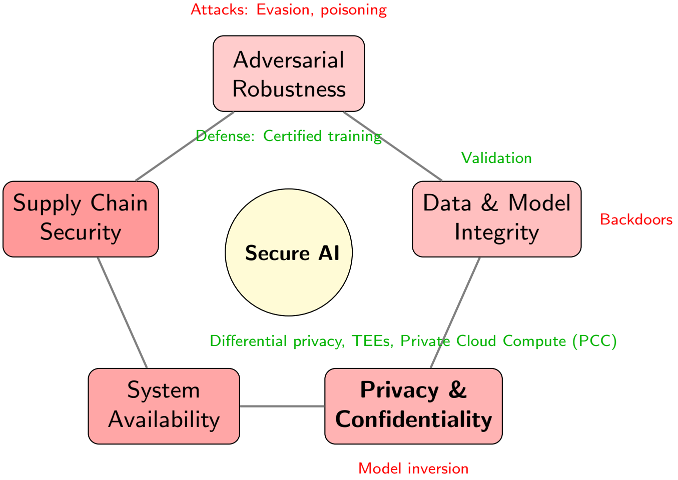
- YouTube: Watch time $\rightarrow$ Radicalization
- Hiring AI: Efficiency $\rightarrow$ Discrimination
- Content moderation: Safety $\rightarrow$ Censorship



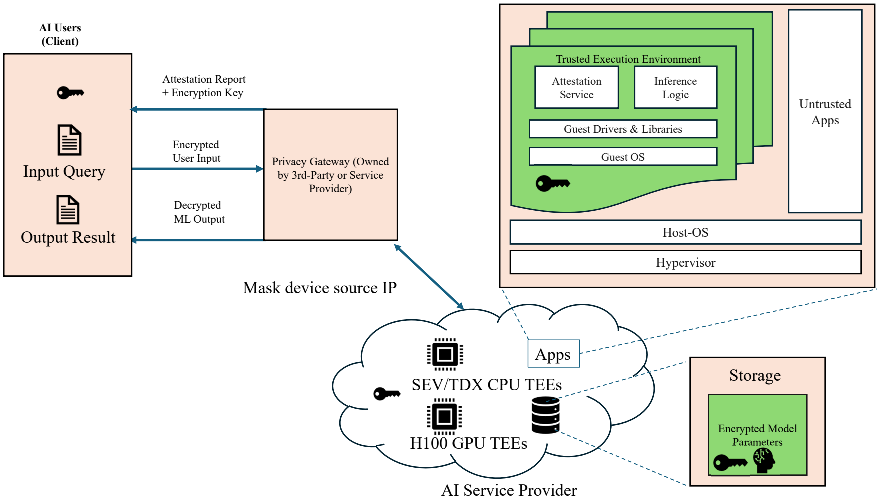
The Stakes

As AI systems become more powerful, alignment failures become more consequential

AI Security Research: Five Domains



Our Ongoing Effort of Securing AI Inferences with TEEs



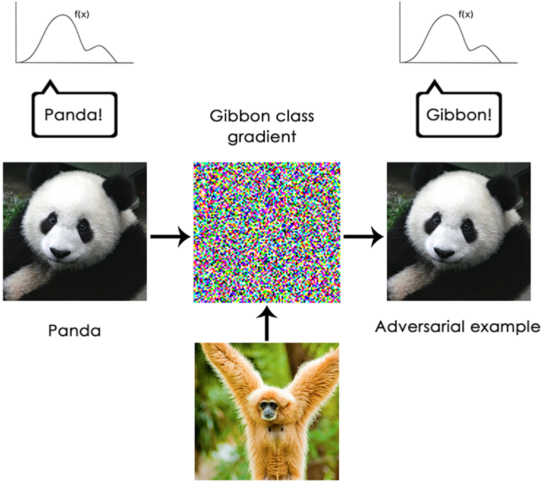
The Arms Race in AI Security

Attack Types

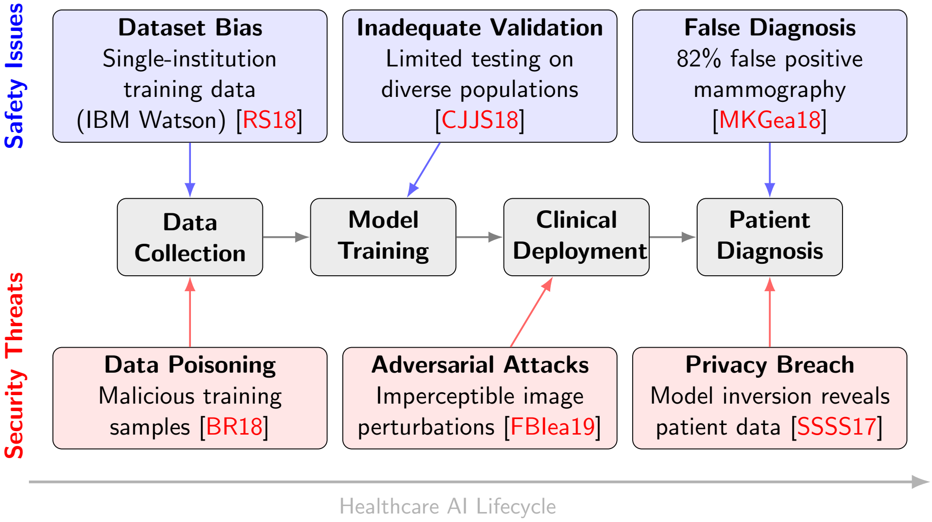
- **Evasion:** Fool deployed models
- **Poisoning:** Corrupt training data
- **Extraction:** Steal model parameters
- **Inference:** Extract private data

Defense Strategies

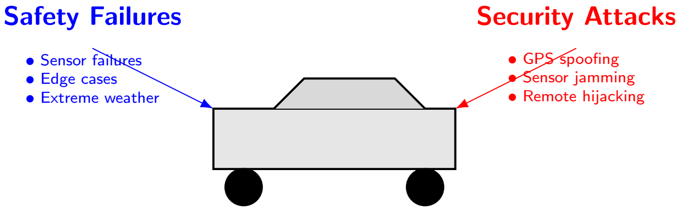
- Adversarial training
- Certified robustness
- Input preprocessing
- Ensemble methods



Case Study 1: Life-Critical Healthcare AI



Case Study 2: Autonomous Vehicles



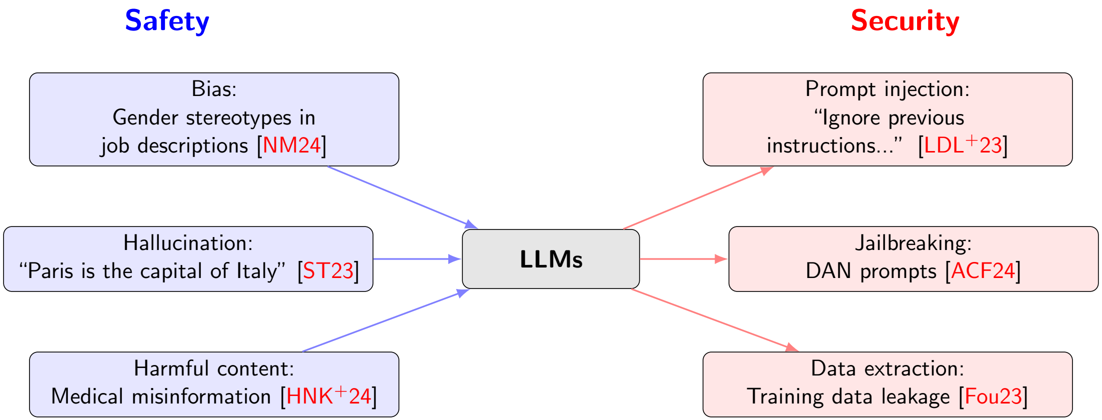
Uber Fatality (2018) - Safety [Dom18]

- Pedestrian detection failure
- Emergency braking disabled
- Human safety driver distracted
- *Solution:* Enhanced sensor fusion, fail-safe mechanisms

Jeep Hack (2015) - Security [Gre15]

- Remote control via internet
- Steering and brakes compromised
- 1.4 million vehicles recalled
- *Solution:* Network isolation, secure update mechanisms

Case Study 3: The Complexity of Generative AI—Large Language Models



AI Safety & AI Security: Different Problems, Different Solutions

AI Safety Research

1. Value alignment [Gab20b]
2. Interpretability (XAI) [GSC+19]
3. Distributional robustness [HZB+19]
4. Bias detection/mitigation [MMS+21]
5. Fail-safe mechanisms [OA16]

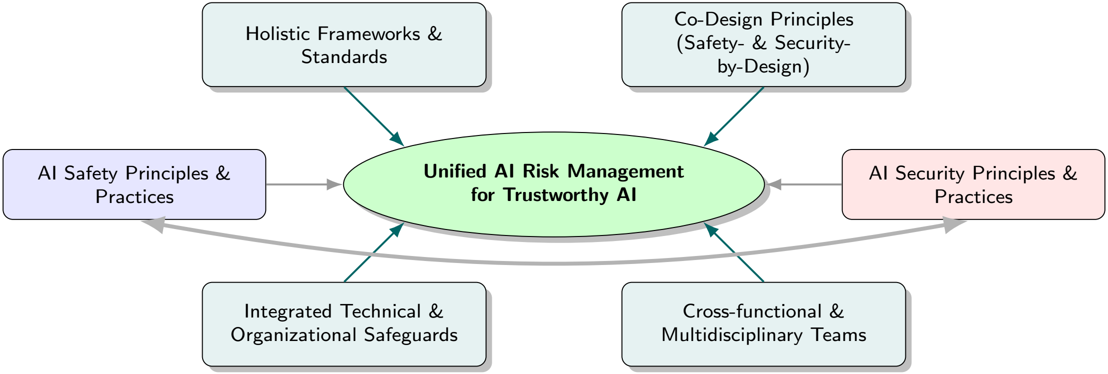
Tools: RLHF [OWJ+22], Constitutional AI [BKK+22], LIME [RSG16], SHAP [LL17]

AI Security Research

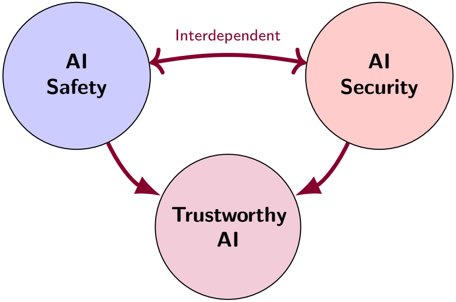
1. Adversarial robustness [MMS+18]
2. Privacy preservation [SSSS17]
3. Model watermarking [UNSS17]
4. Attack detection [AAF+23]
5. Access control [Nat20, BAW+20]

Tools: Adversarial training, Differential privacy, Secure enclaves [SSD22]

The Path Forward: Towards Unified AI Risk Management



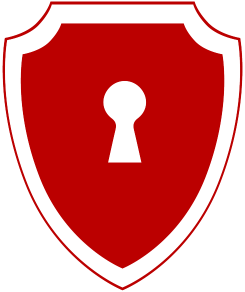
The Path Forward: Towards Unified AI Risk Management



Safe by Design & Secure by Default

About SecLab

About SecLab



Key Research Thrusts

1. (**Why**) Understanding and discovering of **known** or new-emerging (**unknown**) vulnerabilities/attacks/malware
2. (**How**) Developing algorithms, abstractions, (automated) systems, and tools for analysis and defenses

About SecLab

Key Research Thrusts

1. (**Why**) Understanding and discovering of **known** or new-emerging (**unknown**) vulnerabilities/attacks/malware
2. (**How**) Developing algorithms, abstractions, (automated) systems, and tools for analysis and defenses

Current Interests

1. Defense: Systems security (e.g., **TEE/MPC/FHE**, hardening)
2. Offense: Software security (e.g., **reverse engineering**, and **vulnerability** discovery)
3. Security in emerging platforms (e.g., **AI/LLM**, **Agentic AI**, **5G/Satellite**, **blockchain**).

Thank You

Questions & Discussion

zlin@cse.ohio-state.edu

Z. Lin, H. Sun, and N. Shroff. "*AI Safety vs. AI Security: Demystifying the Distinction and Boundaries*". https://www.arxiv.org/abs/2506.18932, June 2025.

References I

- Giovanni Apruzzese, Mauro Andreolini, Luca Ferretti, Mirco Marchetti, and Michele Colajanni, *The role of deep learning in cybersecurity intrusion detection: A comprehensive survey and future challenges*, Journal of Network and Computer Applications **209** (2023), 103540.
- Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion, *Jailbreaking leading safety-aligned llms with simple adaptive attacks*, arXiv preprint arXiv:2404.02151 (2024).
- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané, *Concrete problems in ai safety*, arXiv preprint (2016).
- Tamar Bran et al., *Ai tools in chemical weapons proliferation*, 2023.
- Yoshua Bengio et al., *Managing extreme ai risks in foundation models*, Science (2024).
- Yoshua Bengio et al., *International AI safety report: The international scientific report on the safety of advanced AI*, Tech. report, Produced with support from the UK Government, for the AI Safety Summit initiatives, January 2025.
- Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung, *Toward trustworthy ai development: Mechanisms for supporting verifiable claims*, 2020.

References II

Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and machine learning*, 2019, Online textbook.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al., *Constitutional ai: Harmlessness from ai feedback*, arXiv preprint arXiv:2212.08073 (2022).

Nick Bostrom, *Superintelligence: Paths, dangers, strategies*, Oxford University Press, 2014.

Battista Biggio and Fabio Roli, *Wild patterns: Ten years after the rise of adversarial machine learning*, Pattern Recognition **84** (2018), 317–331.

Irene Y. Chen, Fredrik D. Johansson, Shalmali Joshi, and David Sontag, *Why is my classifier discriminatory?*, NeurIPS, 2018, pp. 3539–3550.

Camila Domonoske, *Ntsb: Uber self-driving car had disabled emergency brake system before fatal crash*, NPR (2018).

Tony Fang et al., *Ai-enhanced cyber capabilities: Capabilities and mitigations*, 2024.

Samuel G. Finlayson, John D. Bowers, Joichi Ito, and et al., *Adversarial attacks against medical deep learning systems*, Science **363** (2019), no. 6433, 1287–1289.

OWASP Foundation, *Llm02:2023 - data leakage*, 2023.

Iason Gabriel, *Artificial intelligence, values and alignment*, Minds and Machines **30** (2020), no. 3, 411–437.

Iason Gabriel, *Artificial intelligence, values, and alignment*, Minds and Machines **30** (2020), 411–437.

References III

- Andy Greenberg, *Hackers remotely kill a jeep on the highway—with me in it*, WIRED (2015).
- David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang, *XAI—explainable artificial intelligence*, Science Robotics **4** (2019), no. 37.
- Tian Han, Sebastian Nebelung, Fadi Khader, et al., *Medical large language models are susceptible to targeted misinformation attacks*, npj Digital Medicine **7** (2024), no. 1, 288.
- Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant, *Risks from learned optimization in advanced machine learning systems*, 2019.
- Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song, *Natural adversarial examples*, arXiv preprint arXiv:1907.07174 (2019).
- Geoffrey Irving, Paul Christiano, and Dario Amodei, *Ai safety via debate*, arXiv preprint arXiv:1805.00899 (2018).
- Hannah Natanson John Hudson, *A marco rubio impostor is using ai voice to call high-level officials - the washington post*, 7 2025.
- Victoria Krakovna, *Specification gaming examples in ai*, 2018.
- Yi Liu, Gelei Deng, Yuekang Li, et al., *Prompt injection attack against llm-integrated applications*, arXiv preprint arXiv:2306.05499 (2023).
- Dave Lee, *Microsoft's tay chatbot returns with 'apology' tweets*, 2016.

References IV

- Scott M. Lundberg and Su-In Lee, *A unified approach to interpreting model predictions*, Advances in Neural Information Processing Systems **30** (2017).
- Liv McMahon, *Ai system resorts to blackmail if told it will be removed*, 2025.
- David Manheim and Scott Garrabrant, *Categorizing variants of goodhart's law*, 2018.
- Diana L. Miglioretti, Karla Kerlikowske, Berta M. Geller, and et al., *Radiologist performance in the national mammography database: Results from 1 million screening mammograms*, Radiology **287** (2018), no. 1, 51–58.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, *Towards deep learning models resistant to adversarial attacks*, 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, *A survey on bias and fairness in machine learning*, ACM Computing Surveys **54** (2021), no. 6, 1–35.
- National Institute of Standards and Technology, *Security and privacy controls for information systems and organizations*, Tech. Report Revision 5, U.S. Department of Commerce, September 2020.
- Guilherme Nomelini and Carla Marcolin, *Gender bias in large language models: A job postings analysis*, RAM. Revista de Administração Mackenzie **25** (2024).

References V

- Laurent Orseau and Stuart Armstrong, *Safely interruptible agents*, Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI) (2016), 557–566.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al., *Training language models to follow instructions with human feedback*, Advances in neural information processing systems **35** (2022), 27730–27744.
- Casey Ross and Ike Swetlitz, *Ibm's watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show*, STAT News (2018).
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, *"why should i trust you?": Explaining the predictions of any classifier*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), 1135–1144.
- Stuart Russell, *Research priorities for robust and beneficial artificial intelligence*, AI Magazine **36** (2015), no. 4, 105–114.
- Karen Scarfone, Murugiah Souppaya, and Donna Dodson, *Secure software development framework (ssdf) version 1.1: Recommendations for mitigating the risk of software vulnerabilities*, Special Publication (NIST SP) 800-218, National Institute of Standards and Technology, Gaithersburg, MD, February 2022.
- Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, *Membership inference attacks against machine learning models*, Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), IEEE, May 2017, pp. 3–18.
- Marco Siino and Ilenia Tinnirello, *Gpt hallucination detection through prompt engineering*, Working Notes of CLEF 2024, CEUR Workshop Proceedings, vol. 3740, 2023, p. 69.

References VI

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh, *Embedding watermarks into deep neural networks*, Proceedings of the International Conference on Machine Learning (ICML) Workshop on Reproducibility in Machine Learning, 2017.

Daisuke Wakabayashi, *Self-driving uber car kills pedestrian in arizona, where robots roam*, 2018.