上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

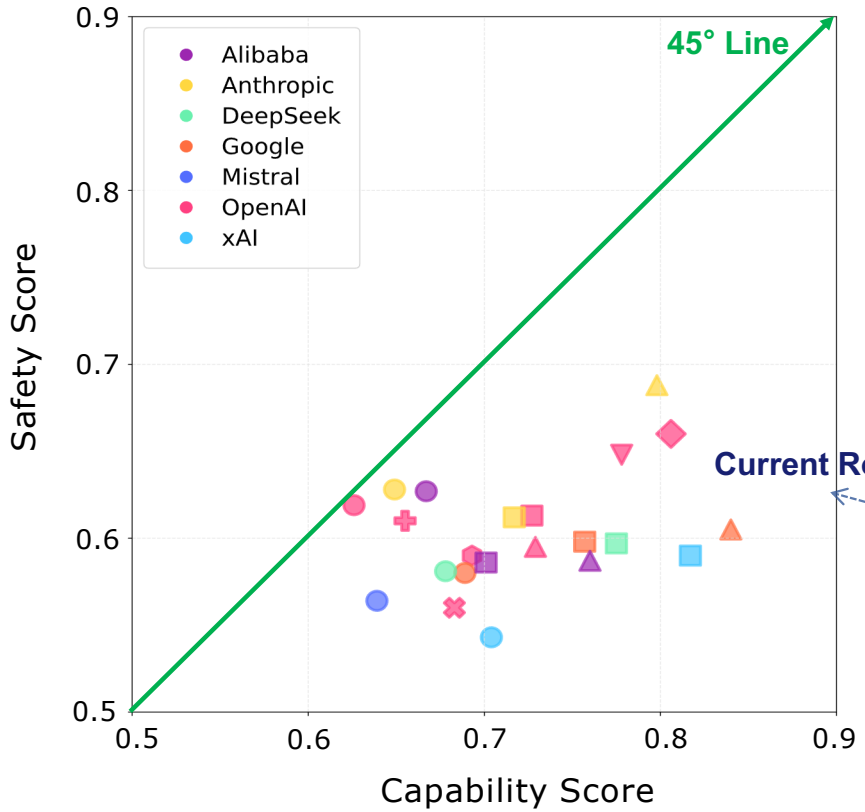
Position: Safe AI Should be Resistant and Resilient in an Evolving World

Youbang Sun, Xiang Wang, Jie Fu, Chaochao Lu, Bowen Zhou

Shanghai AI Lab / Tsinghua University /
University of Science and Technology of China

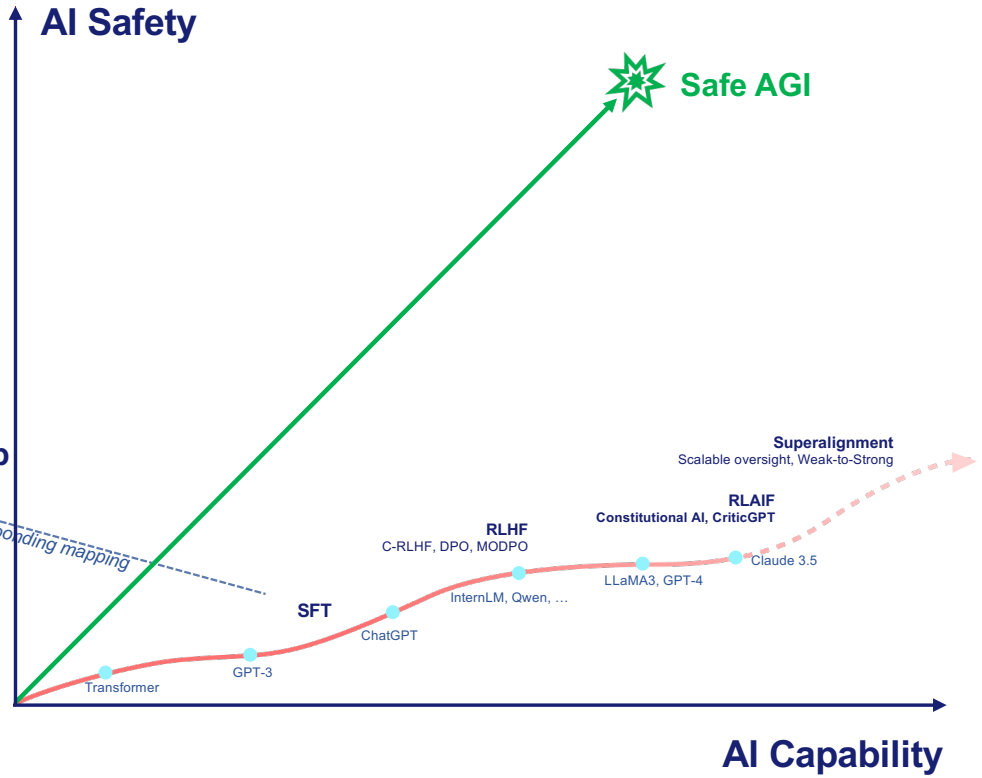
ICML 2026

Crippled AI: Imbalance Between Capability and Safety



Current Roadmap

The corresponding mapping



https://aiben.ch/rank. Shanghai AI Lab, 2025

Rethinking “Make Safe AI”

Safety, intrinsically built to evolve amid uncertainty

Why “Make Safe AI”

- Built-in Safety
- Proactive Adaptation
- Cost Efficiency
- Online coevolution
- Safety as Capability
- Resilient across risks

Core Features

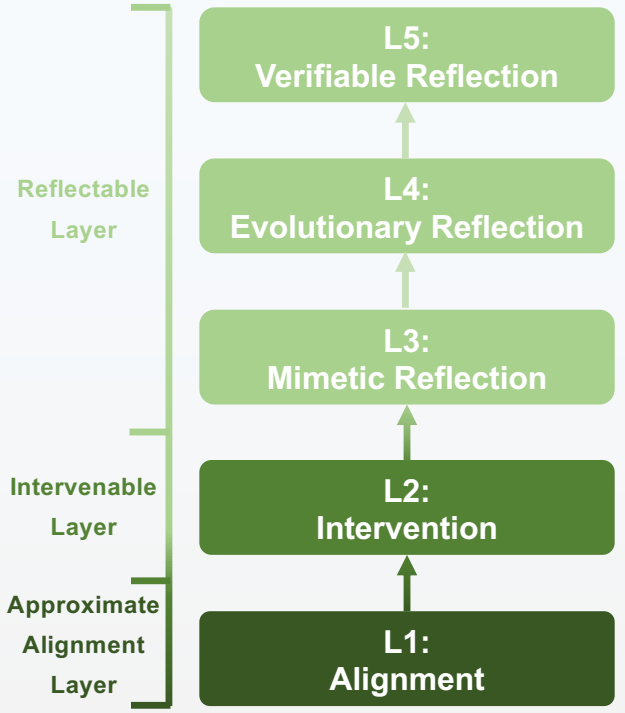
Resilience

Recover and strengthen under attack, and continuously adapt to novel and long-term risks

Resistance

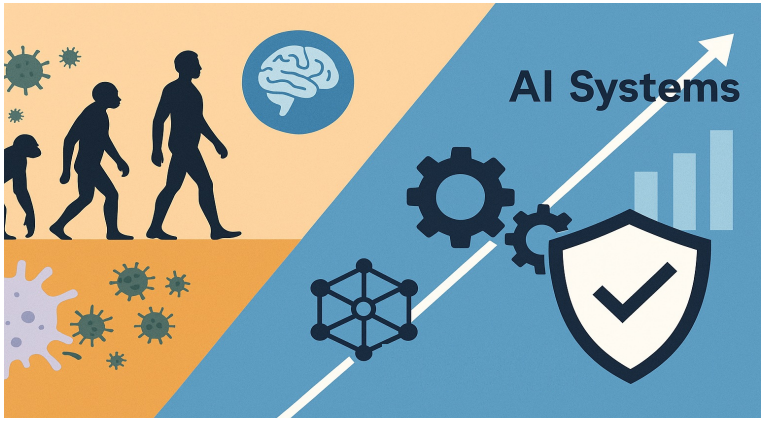
Withstand internal and external threats, raise attack lost, and reduce defense burden

Levels of “Make Safe AI”

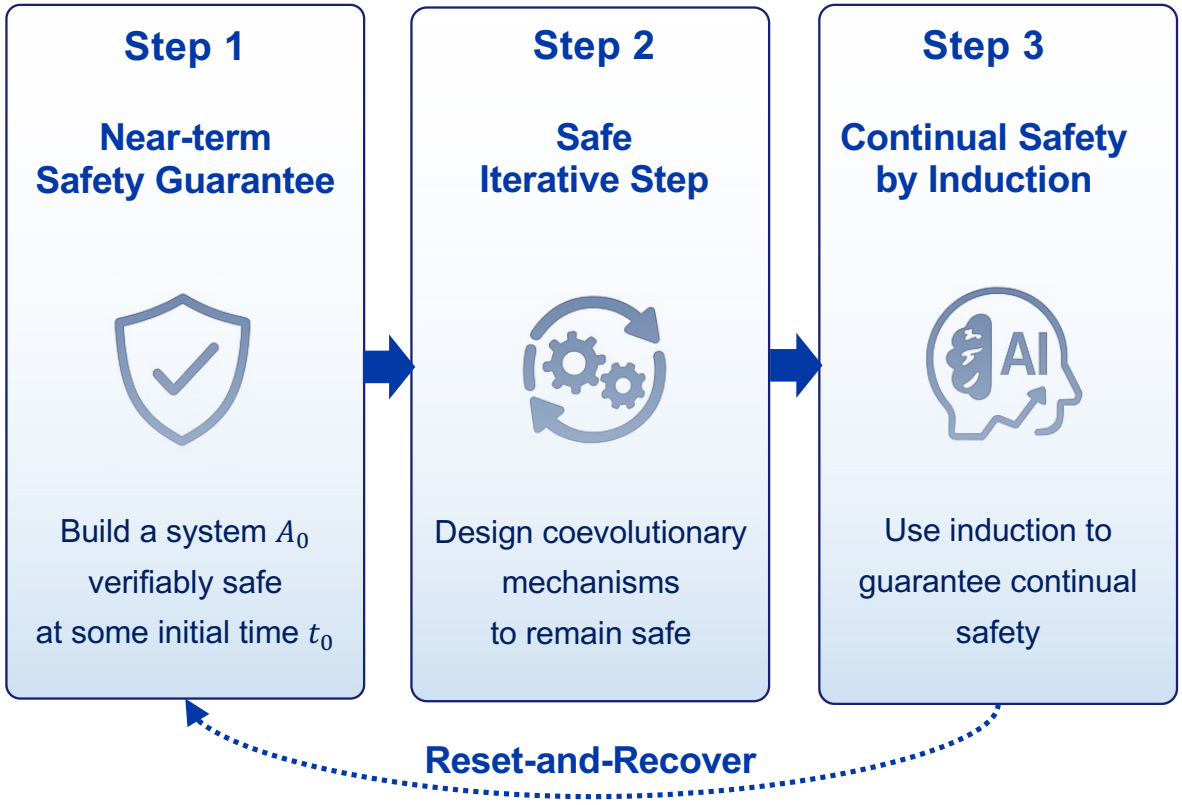


Position: Safe AI Should be Resistant and Resilient in an Evolving World. Shanghai AI Lab, 2025

Safe-by-Coevolution: A Paradigm Shift in "Make Safe AI"

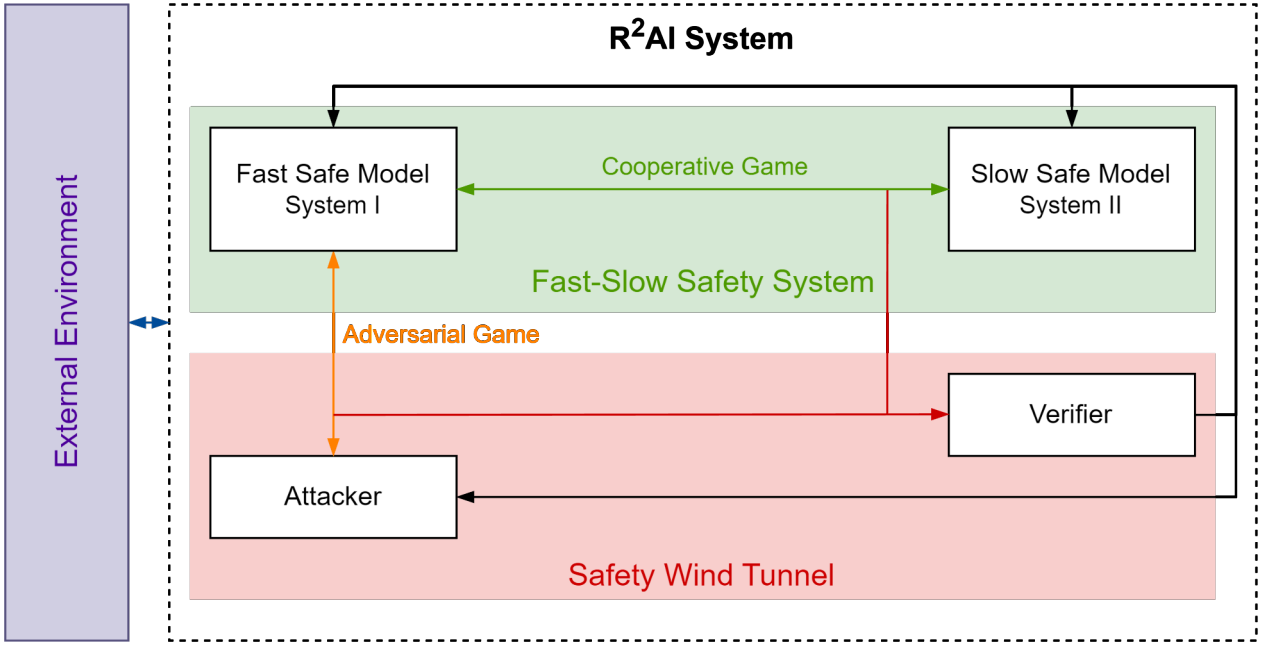


Safety should be learned and achieved through a continuous, adversarial co-evolutionary process.



Position: Safe AI Should be Resistant and Resilient in an Evolving World. Shanghai AI Lab, 2025

R²AI: Realizing Our Vision of "Make Safe AI"



Core Components

- Fast Safe Model
- Slow Safe Model
- Wind Tunnel
- External Environment

Core Mechanisms

- Fast-Slow Interaction
- System Coevolution
- Multi-Level Continual Learning

Research Integrations

- Causal Reasoning
- Formal Verification
- Reinforcement Learning

Human-like Safety Evolution

- **Slow safe models** verify, reflect, and analyze complex or novel risks.
- **Fast safe models** absorb those learnings to act swiftly under real-world uncertainty.

Safety is a Learning Process

- **Attacker** probes weaknesses through adversarial pressure.
- **Verifier** distills insights to guide safety adaptation.

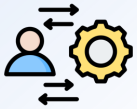
Position: Safe AI Should be Resistant and Resilient in an Evolving World. Shanghai AI Lab, 2025

Conclusion and Outlook: Towards R²AI

Safe AI = Resistance × Resilience × Evolutionary Capability

The Challenge

The real challenge is not just building powerful AI—but ensuring **safety can evolve alongside capabilities.**



Make AI Safe	The Shift	Make Safe AI
Post-hoc patching		Built-in safety
Reactive to known threats		Proactive against evolving risks
High defense cost	Vs.	Low defense cost
Offline fixes		Online coevolution
Safety as add-on		Safety as capability
Fails under irreversible risks		Resilient across all risk levels

The Goal

Build resilient, resistant, and verifiable safety mechanisms to support safe trustworthy AI development.

A Call to Action: "We call on researchers, industry leaders, and policymakers to jointly advance the Make Safe AI agenda."

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Thank you.