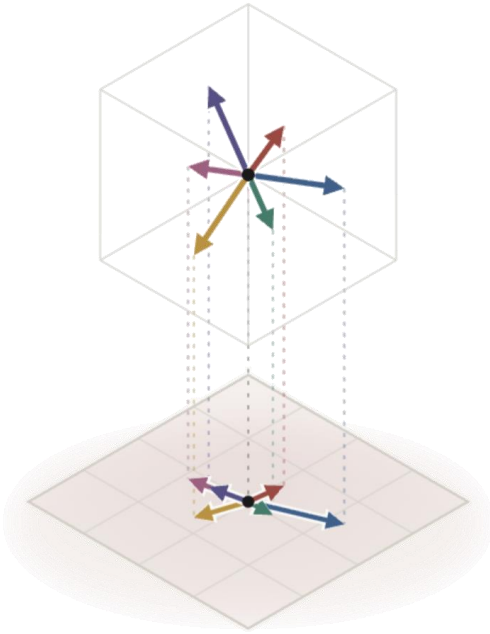
Confirming Claims of Superposition and Adversarial Examples in Image Classifiers

Ray Che

Existential Risk Laboratory at University of Chicago



Safety vs Security

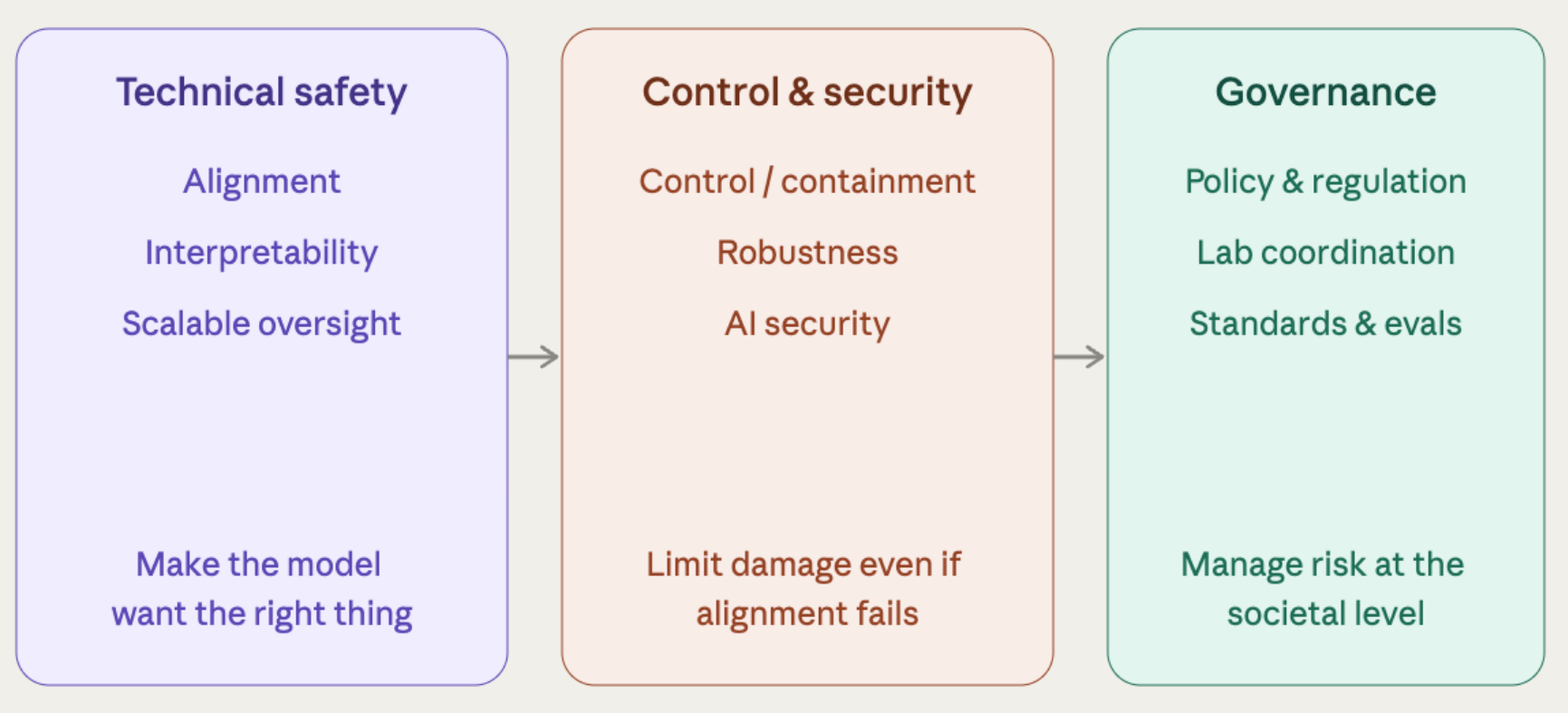
Safety - No adversary.

- Misalignment
- Hallucination
- Reward hacking
- Evaluation awareness

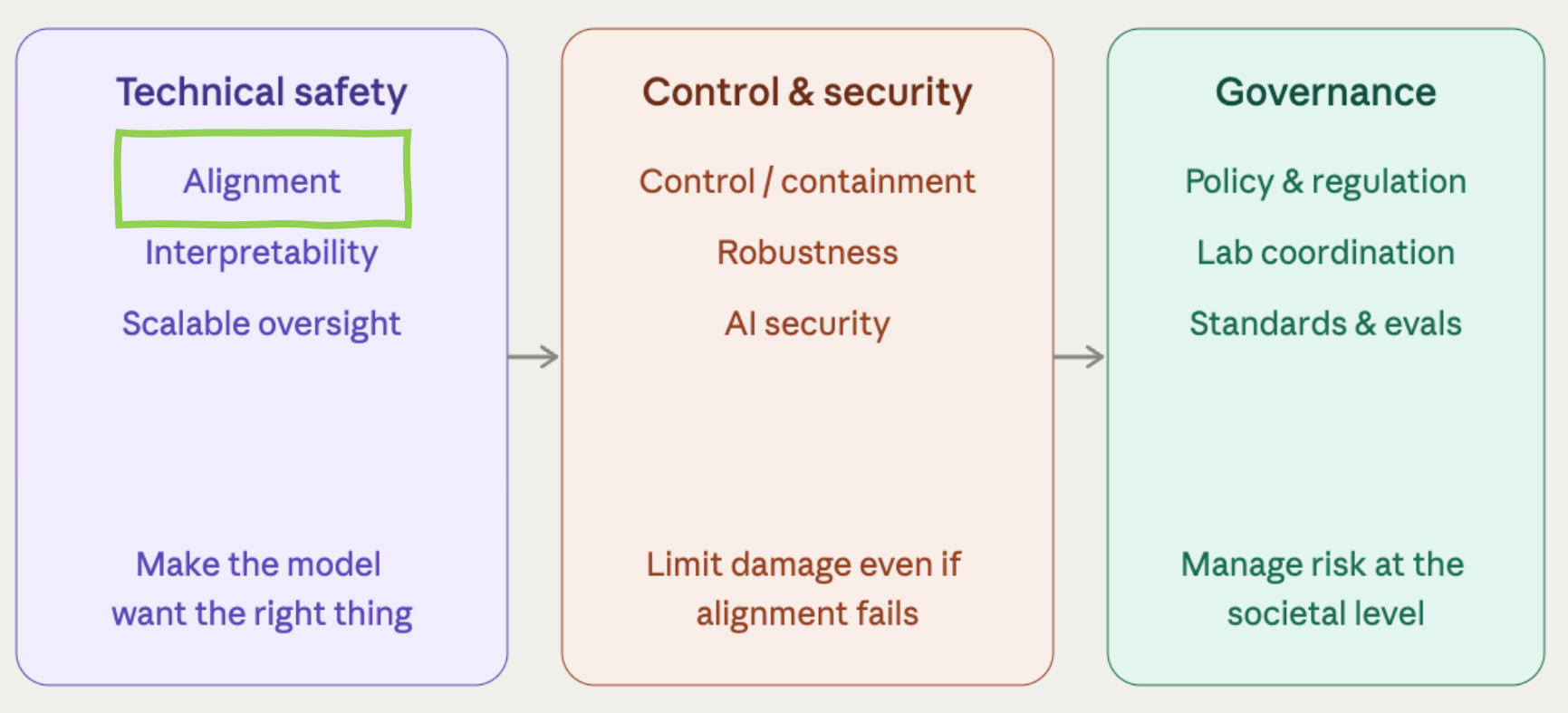
Security - Adversaries present.

- AI
  - Jailbreaking
  - prompt injection
- Traditional
  - Network security
  - System security

AI Safety Research

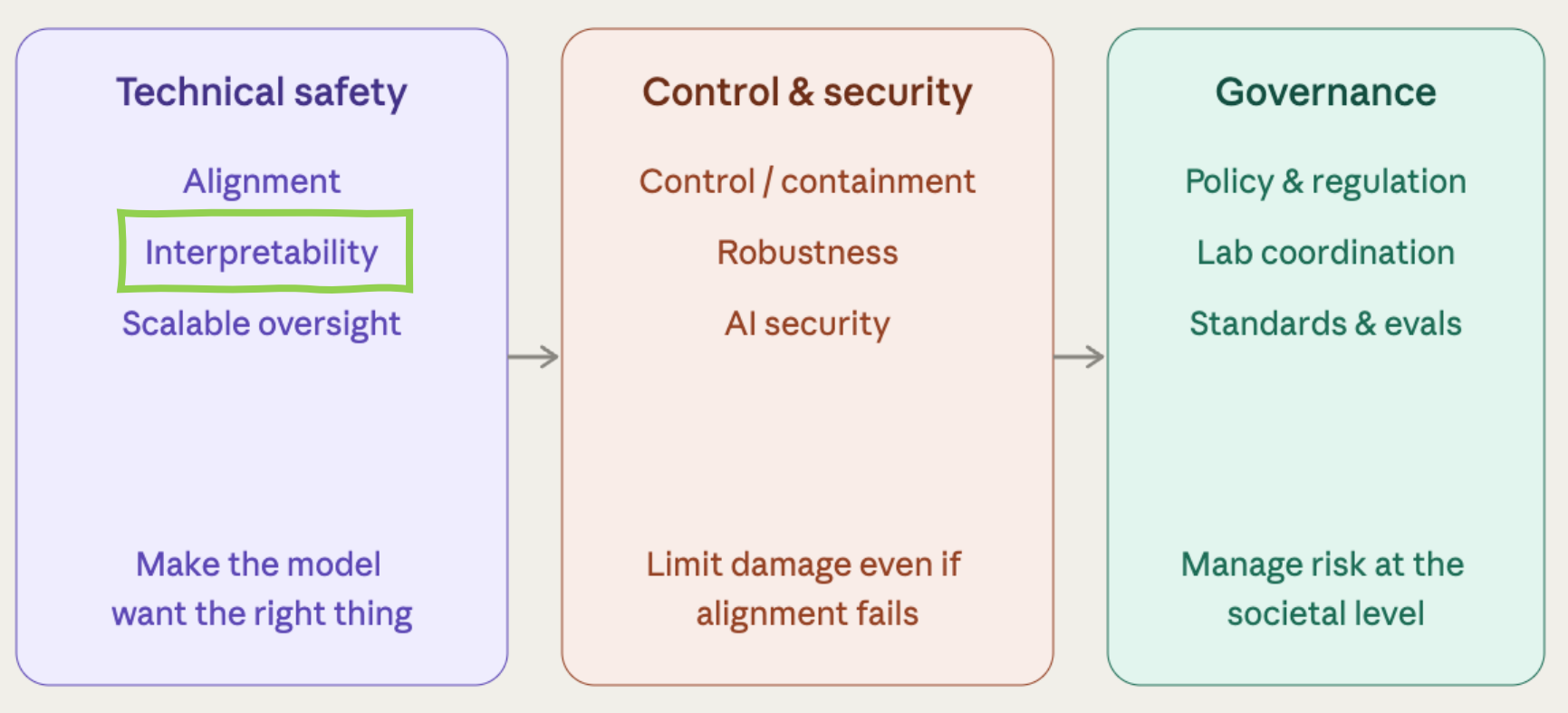


AI Safety Research



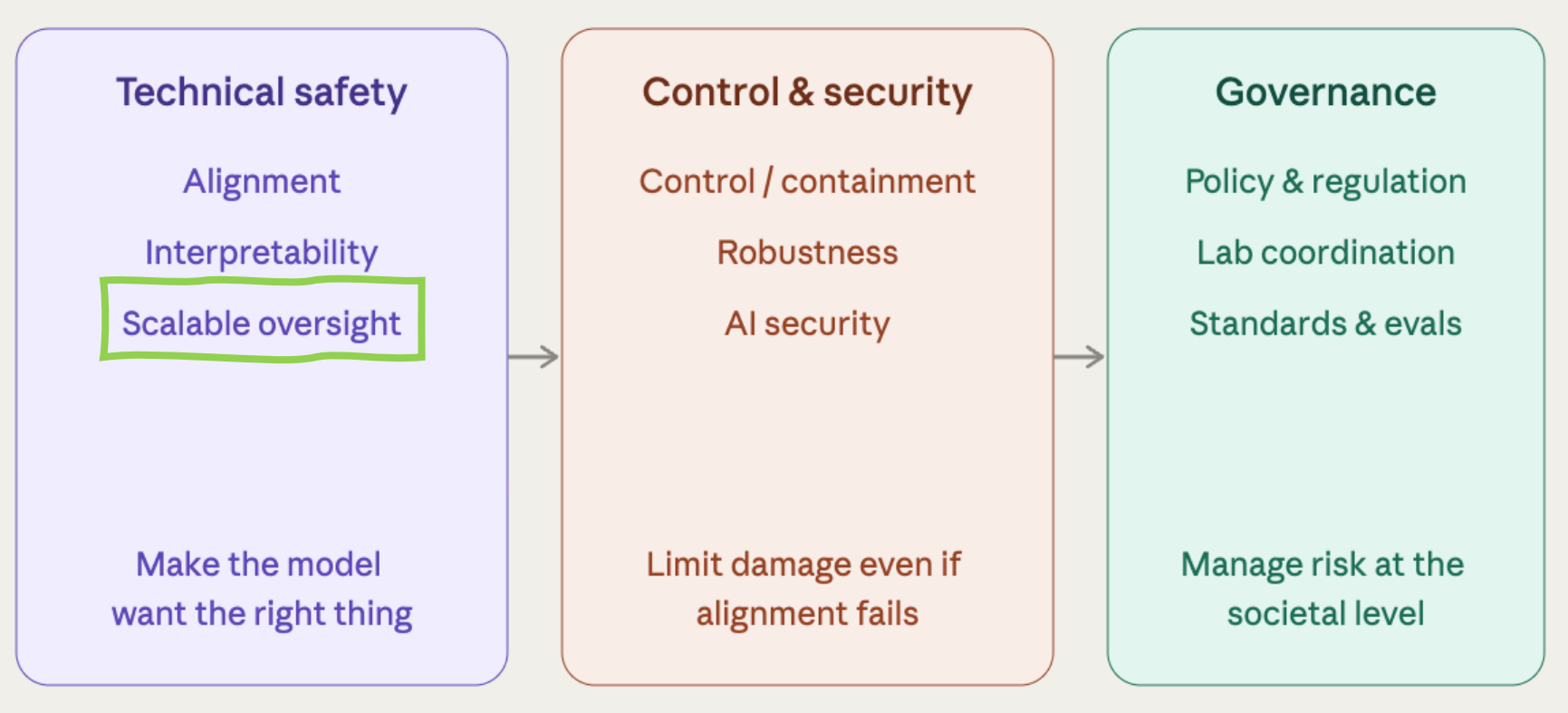
Making sure a model's goals and behavior match what humans actually want, not just what training signals literally reward. Includes techniques like *RLHF*, *Constitutional AI*, and *reward modeling*.

AI Safety Research



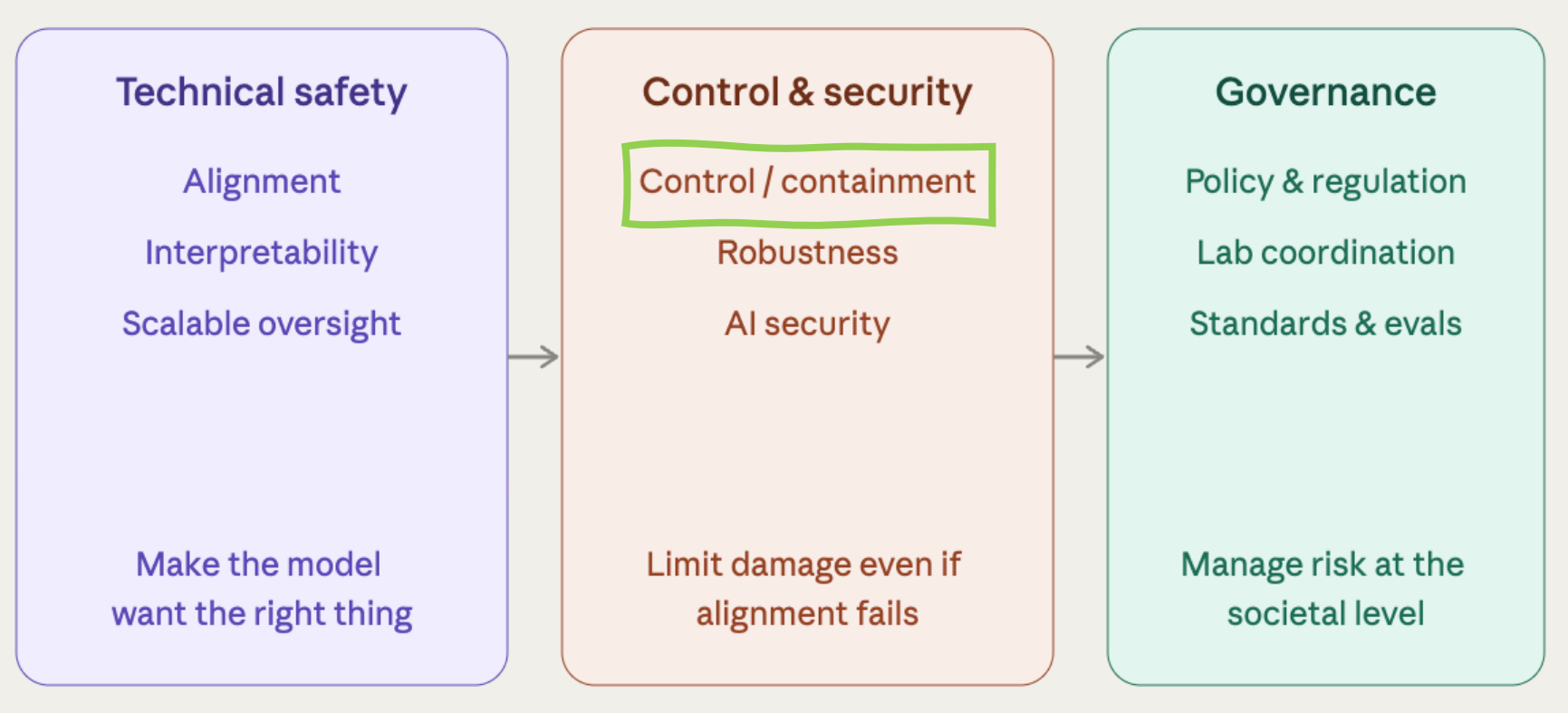
Understanding what's happening inside a model. *Mechanistic interpretability* tries to reverse-engineer internal circuits and representations.

AI Safety Research



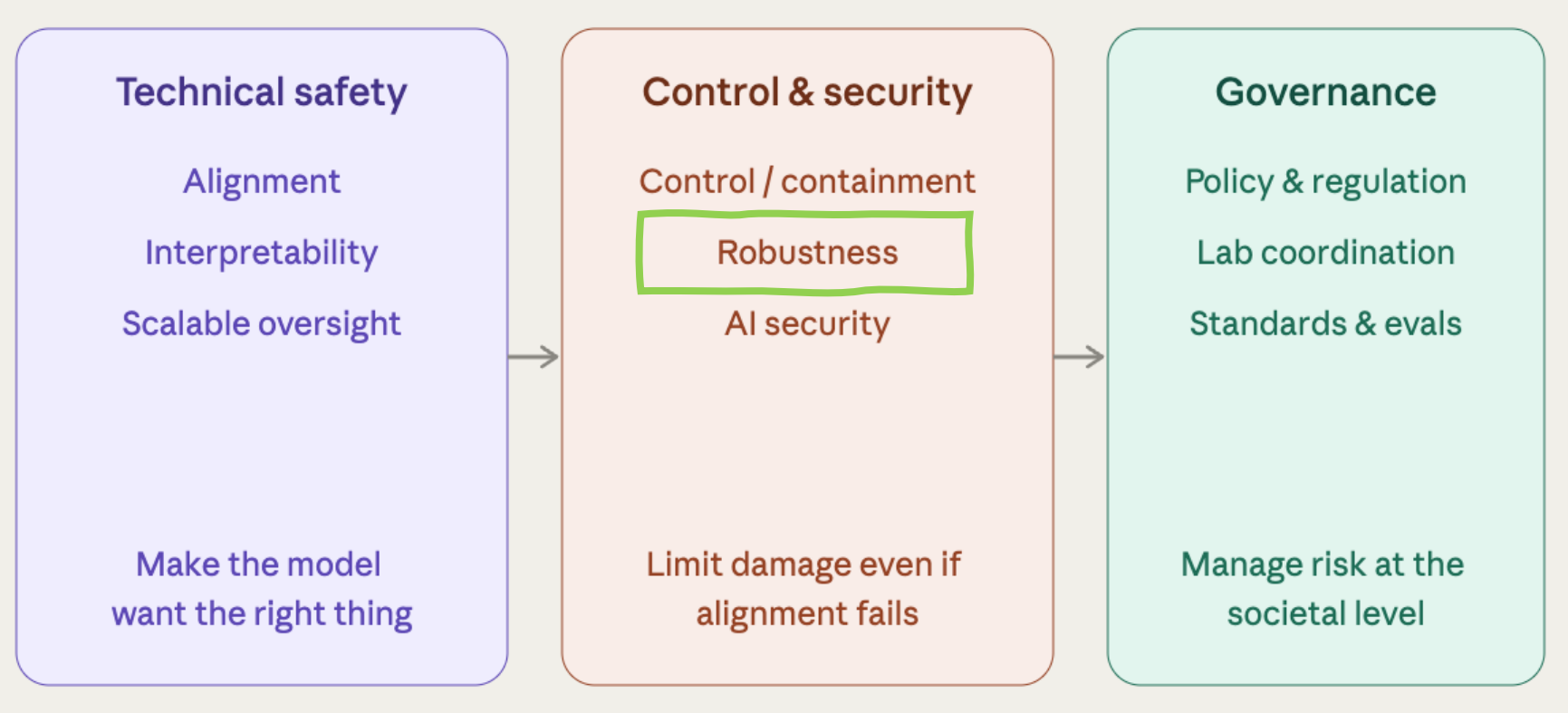
How humans can supervise or evaluate systems that may eventually exceed human ability in a domain.

AI Safety Research



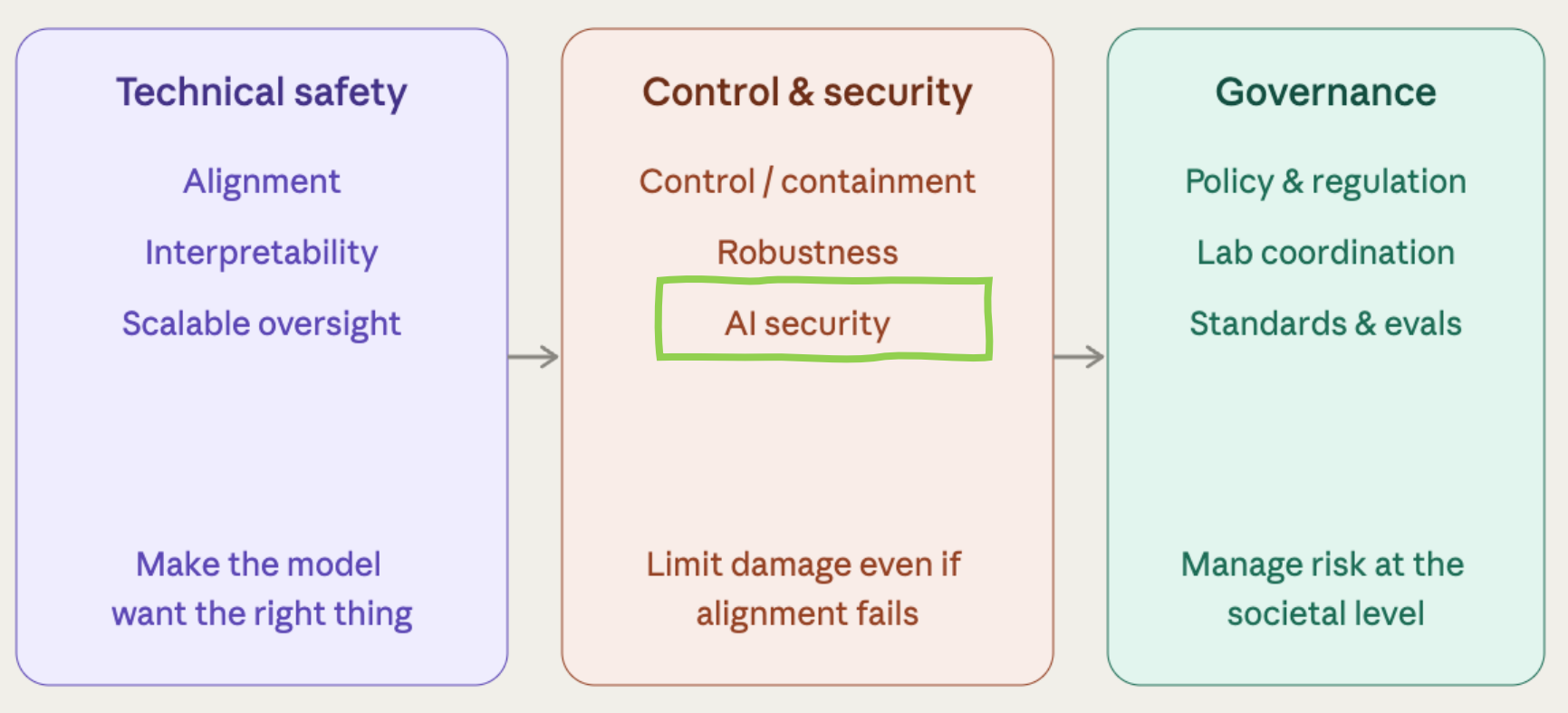
Limiting the damage a misaligned or misused system could do, independent of whether it's "aligned" (*sandboxing*, *monitoring*, *red-teaming*, *evals* for dangerous capabilities).

AI Safety Research



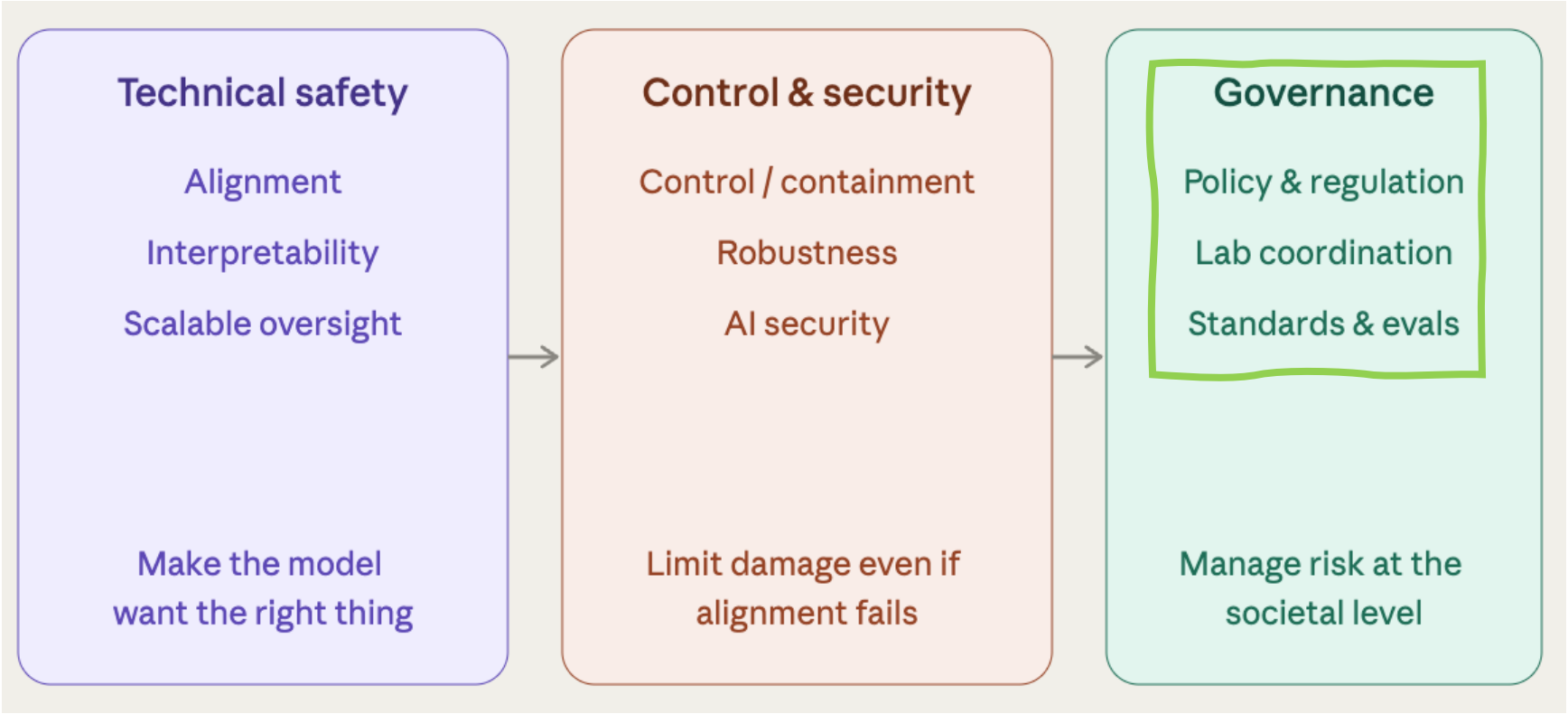
Ensuring models behave reliably under distribution shift, adversarial inputs, or edge cases (e.g., *adversarial examples*, *jailbreak* resistance).

AI Safety Research

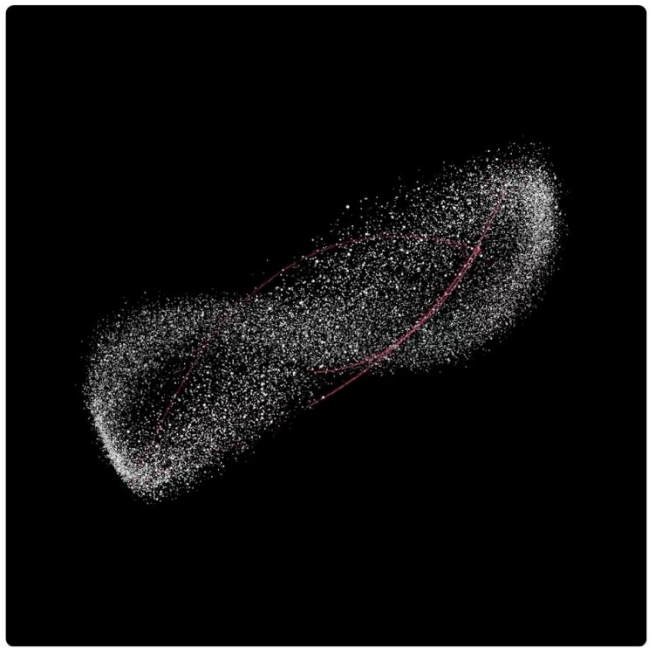


Protecting model *weights* from theft, securing training *infrastructure*, defending against *data poisoning* and *supply-chain attacks*, preventing extraction/inversion attacks that leak training data.

AI Safety Research



Security

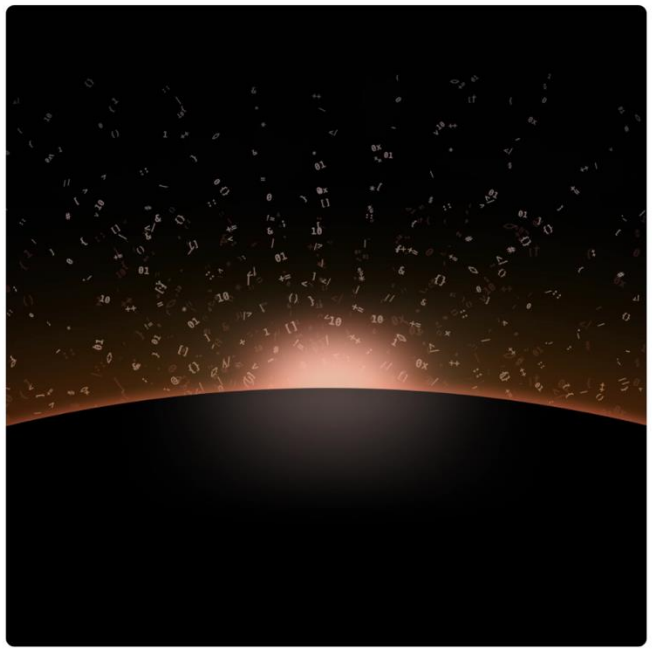


The Hugging Face incident and the road ahead

Security Aug 26, 2026

The Defender's Window

Security Aug 17, 2026



Expanding Daybreak as the Cyber Defense Window Narrows

Security Aug 10, 2026

AI Safety Opportunities

- https://aisopportunities.com
- SPAR projects: https://sparai.org/projects/f26/
- MATS: https://www.matsprogram.org
- LASR: https://www.lasrlabs.org

Things are moving at a fast pace!

Anthropic Fellows Program for AI safety research: applications open for November 2026

"This is an exceptional opportunity to join AI safety research, collaborating with leading researchers on one of the world's most pressing problems." — Jan Leike

July 20, 2026

The Anthropic Fellows program provides funding and Anthropic mentorship for engineers and researchers to investigate some of Anthropic's highest priority AI safety research questions.

April 6, 2026 Safety

Introducing the OpenAI Safety Fellowship

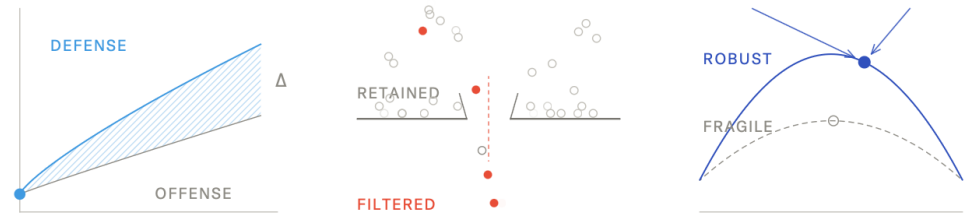
A pilot program to support independent safety and alignment research and develop the next generation of talent

Apply now ↗

Announcing Safety Research Grants

Thinking Machines

Aug 24, 2026



We recently released Inkling and Inkling-Small with open weights, and explained how we are thinking about safety and openness. Today, we are launching Tinker grants of up to $50,000 in credits for safety research on open-weight models. Below are some directions we find promising. The list is non-exhaustive on purpose. If you're working on a safety project that could be accelerated by additional Tinker credits, we want to hear from you.

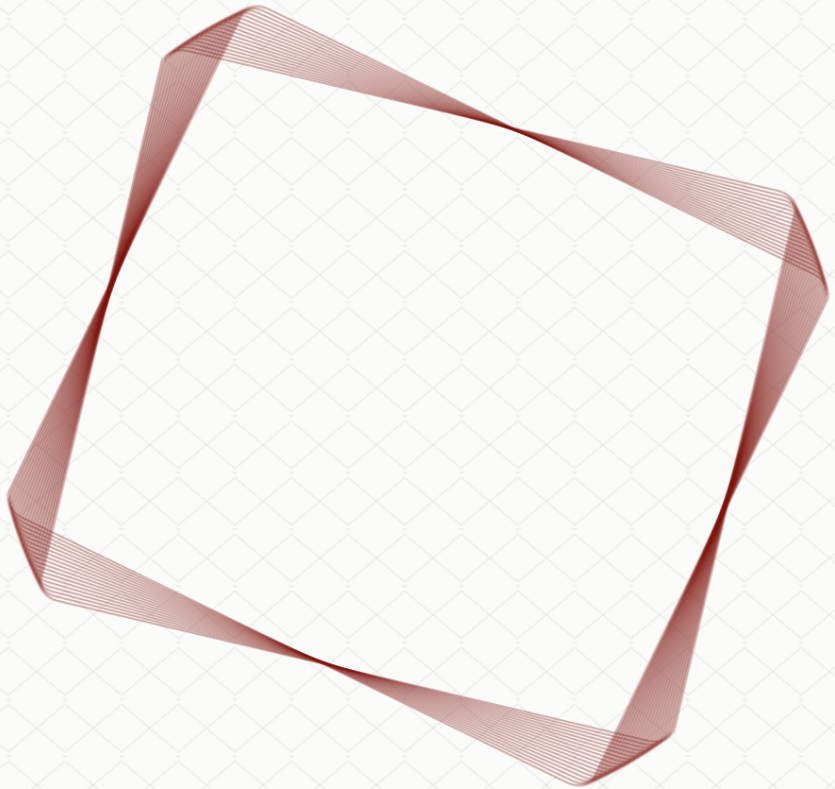
Second Look Research

Replicating Empirical AI Safety Research

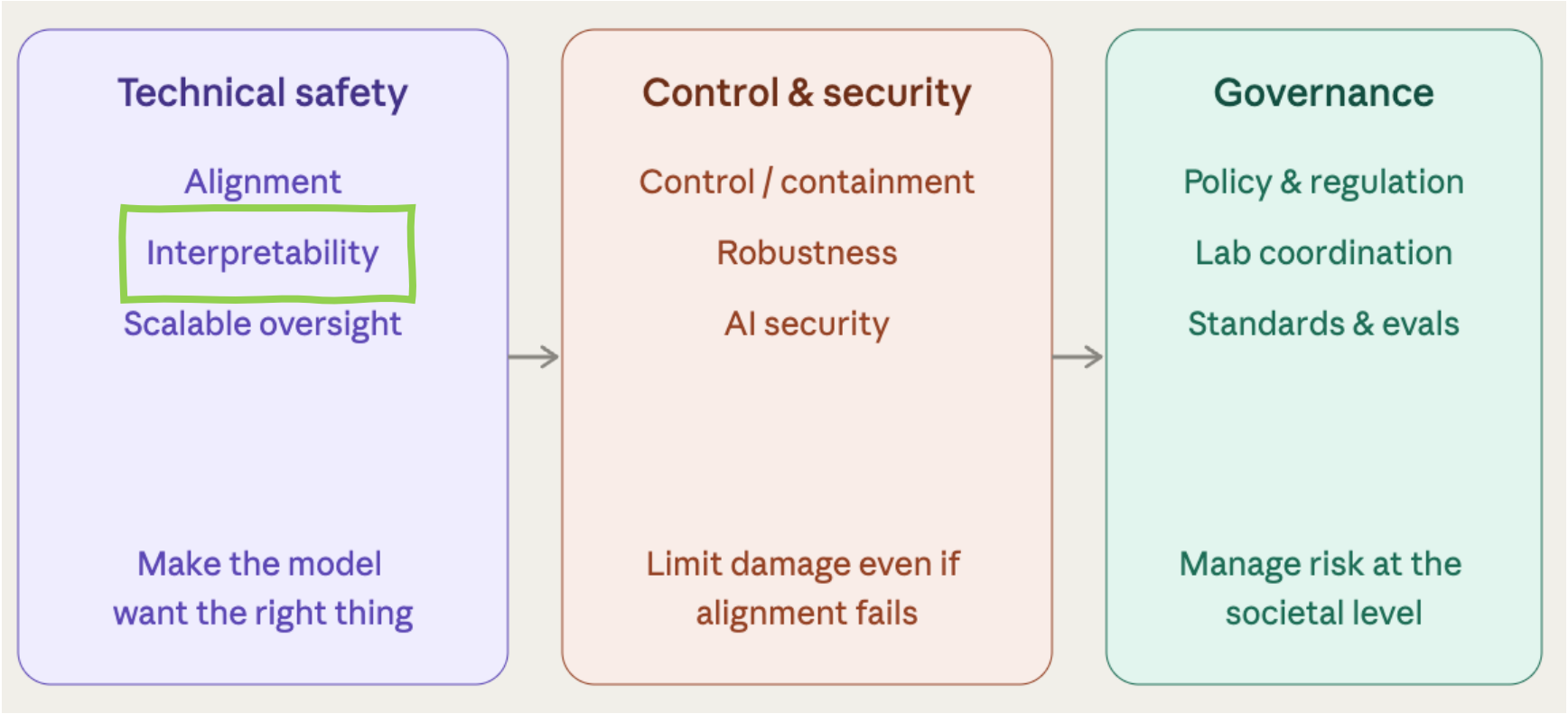
Second Look Research is a project from the **Existential Risk Lab at the University of Chicago**. We produce open source replications of load-bearing results in AI safety.

VIEW RESEARCH

REQUEST A REPLICATION



AI Safety Research

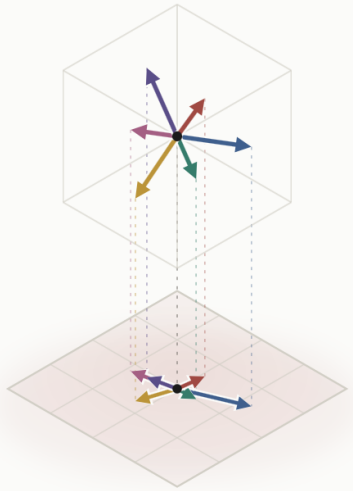


Confirming Claims of Superposition and Adversarial Examples in Toy Models

Xijia Che[†] and Zephaniah Roe[‡]

[†]Corresponding author [‡]Advisor

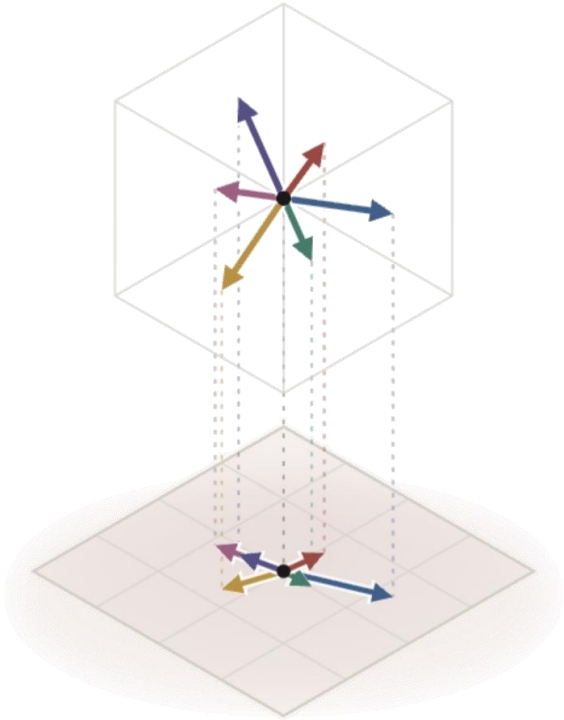
July 27, 2026



Superposition: when a model has fewer dimensions than the data has features, it cannot give each feature its own direction — projecting them into the bottleneck forces them to share axes and interfere.

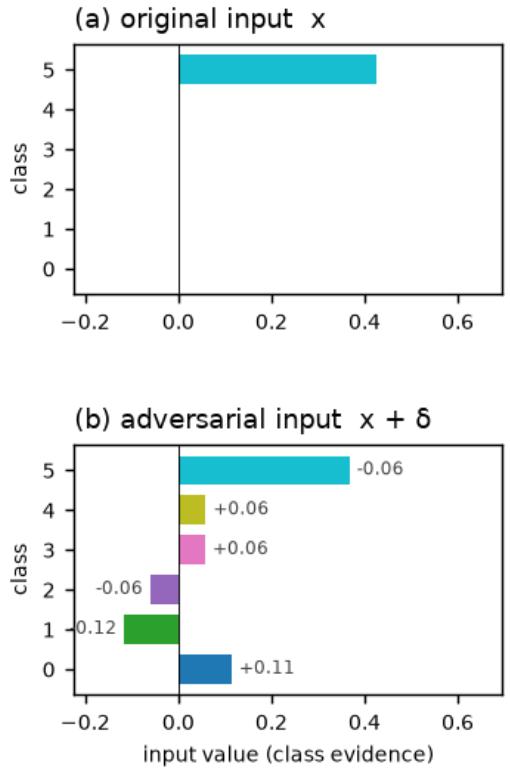
Superposition

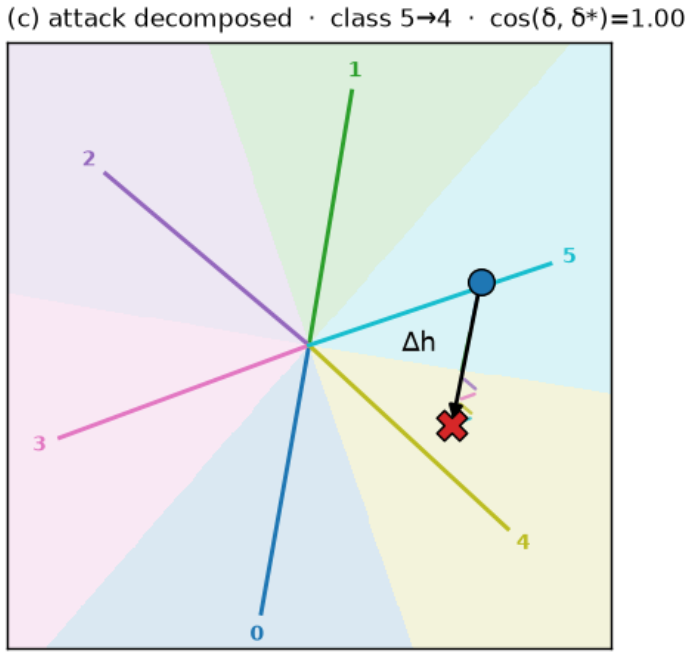
- the phenomenon where “models represent more features than they have dimensions.”

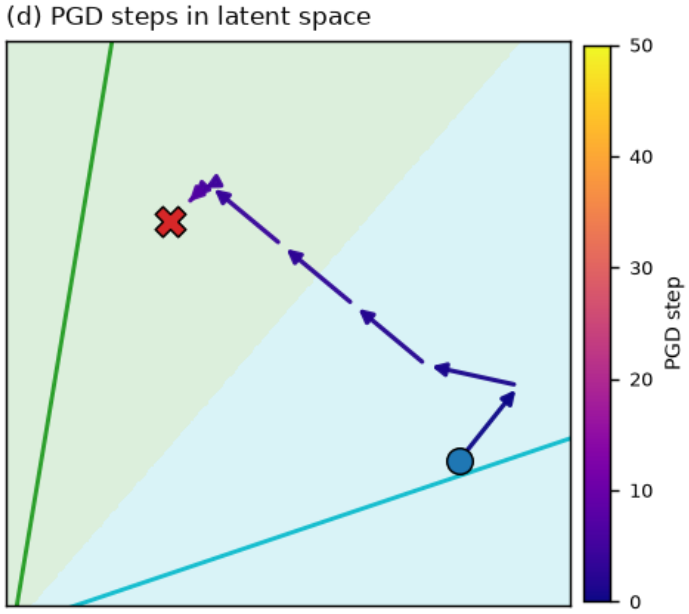


Is superposition the reason for adversarial attacks?

PGD attack


(a) original input x
class
5
4
3
2
1
0
−0.2
0.0
0.2
0.4
0.6
(b) adversarial input x + δ
-0.06
+0.06
+0.06
-0.06
0.12
+0.11
input value (class evidence)


(c) attack decomposed · class 5→4 · cos(δ, δ*)=1.00
1
2
3
4
5
0
Δh


(d) PGD steps in latent space
PGD step
50
40
30
20
10
0

Toy model

$$W_e \in \mathbb{R}^{m \times k}$$

$$W_d \in \mathbb{R}^{k \times m}$$

$$m < k$$

$$\delta \propto W_e^\top (w_t - w_s)$$

- Encoder
- Decoder
- Artificial bottleneck: m
- Number of classes: k

Attacks exploit interference

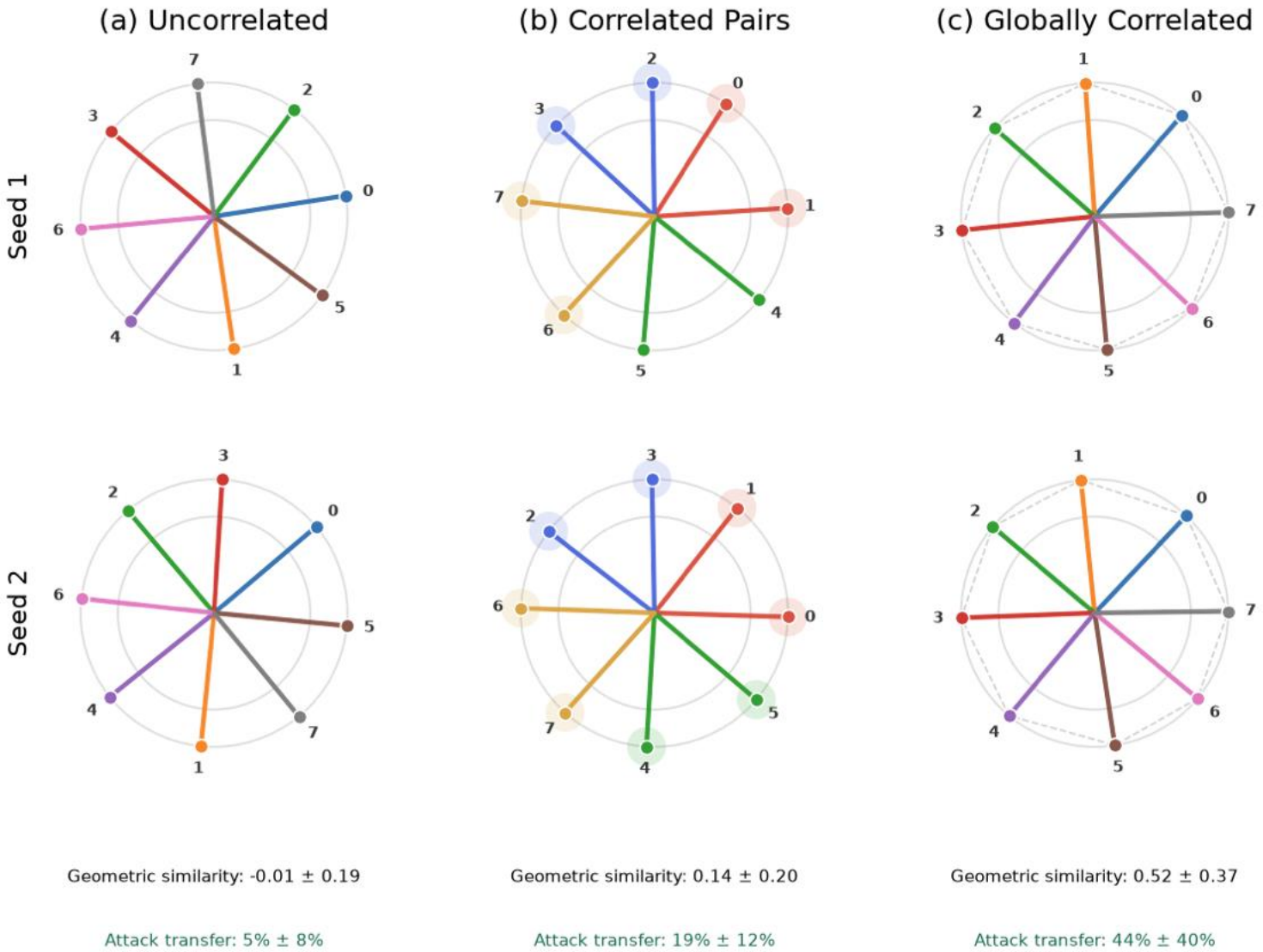
k	m	clean acc	PGD·optimal (ours)	PGD·optimal (paper)
6	2	0.998	1.00	0.97
30	10	0.996	1.00	0.96
90	30	0.966	1.00	0.92

No superposition, little vulnerability

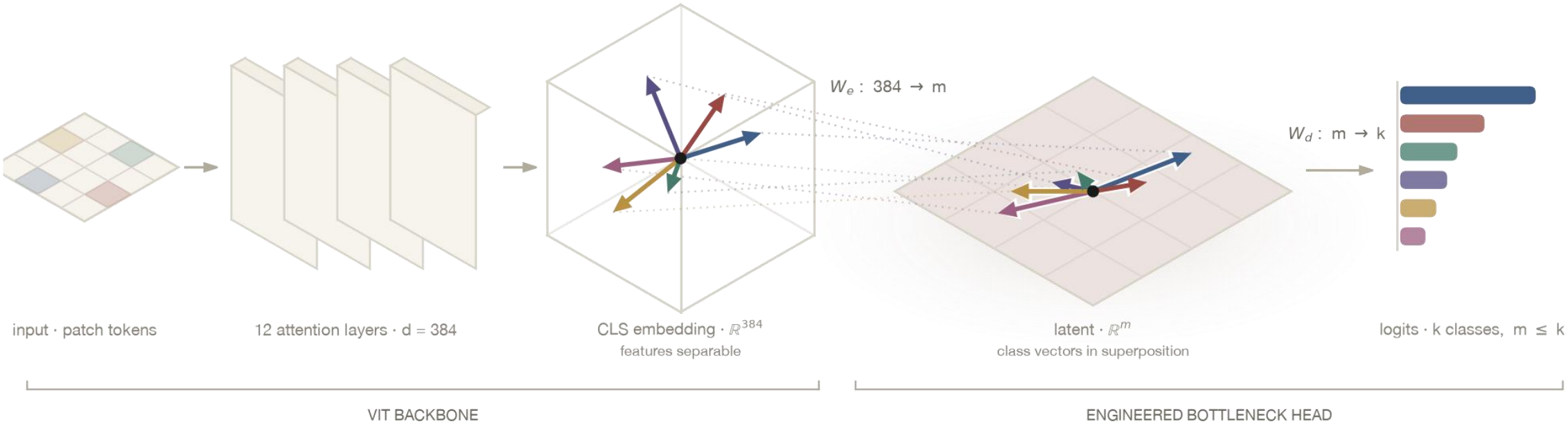
k/m	Paper reported (m=2)	Released code (m=2)	Our replication (m=2)	Paper reported (m=5)	Released code (m=5)	Our replication (m=5)
1	98.7	98.6	100	92.0	95.0	97.3
2	75.8	80.1	79.5	87.2	97.0	70.5
3	31.0	34.3	50.4	59.0	66.4	50.6
4	11.1	11.4	18.7	33.5	36.7	30.1

Robust accuracy vs k/m

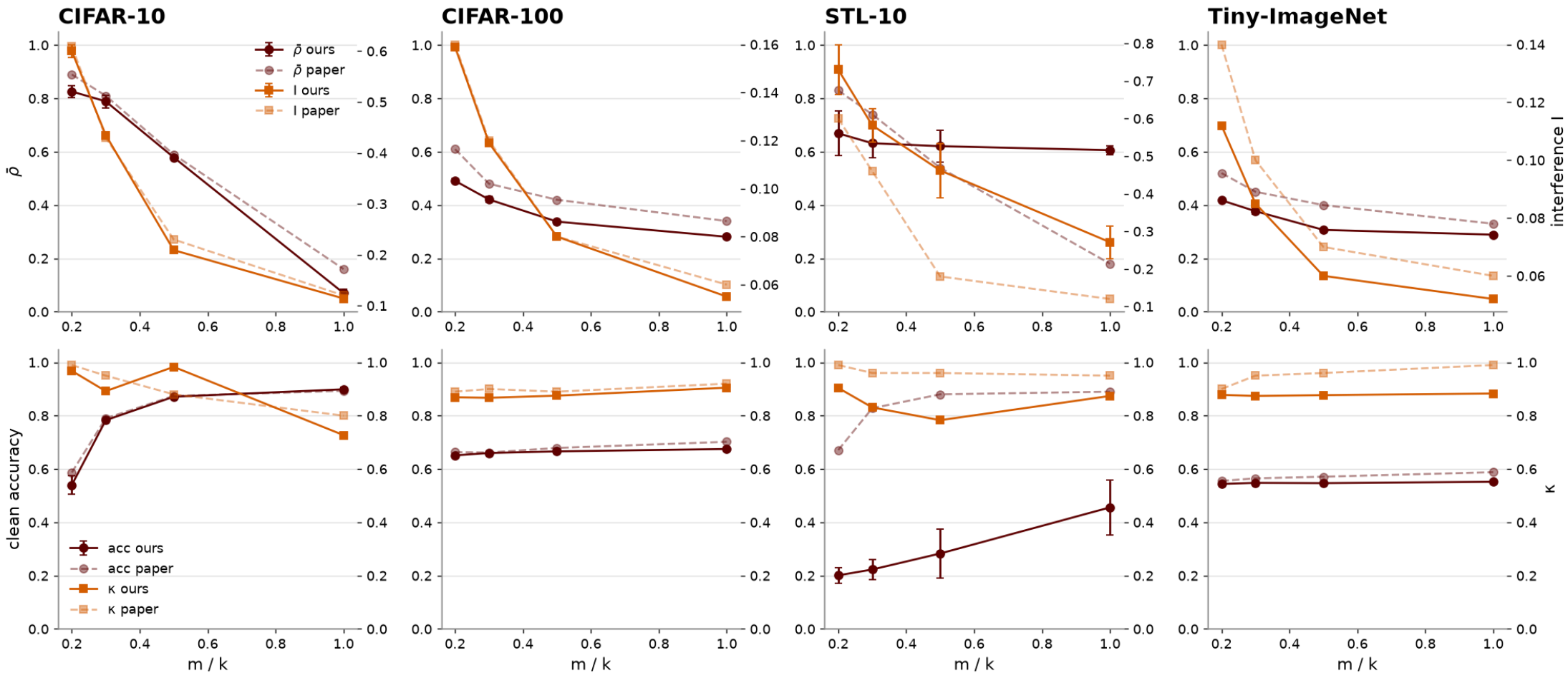
Shared geometry drives transfer attack



Vision Transformer



Interference Predicts Attacks as Bottleneck Tightens



Extensions

- Multiple attack families
  - Coupling Is Attack-independent
- Adversarial training
  - Geometry Is Not The Lever for Robustness
- Larger models
  - Capacity Changes Input and Latent Space Correlation

Conclusion

- Superposition contributes to adversarial attacks.
- There are lots of talents, attention, resources, anxiety, and noise in AI safety.
- The industry is changing at a pace that is way more faster than what I perceive in the academia.

This is a replication of Adversarial Attacks Leverage Interference Between Features in Superposition, completed as part of the Second Look Summer Fellowship.

Supervisor: Zephaniah Roe