AISI | AI SECURITY INSTITUTE

Asymptotic guarantees for alignment

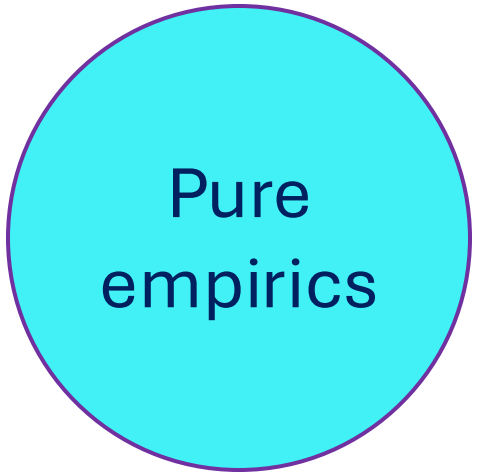
Geoffrey Irving, 15 April 2025

AISI

Please ask questions throughout!

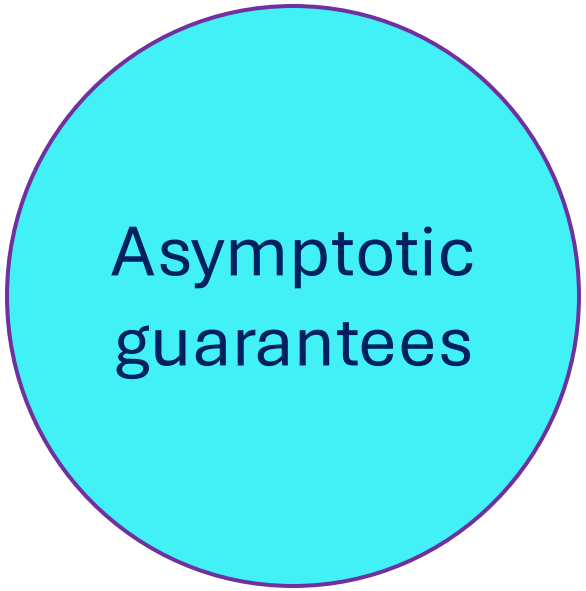
Claim: Theory can tell us what empirics should measure

Too weak



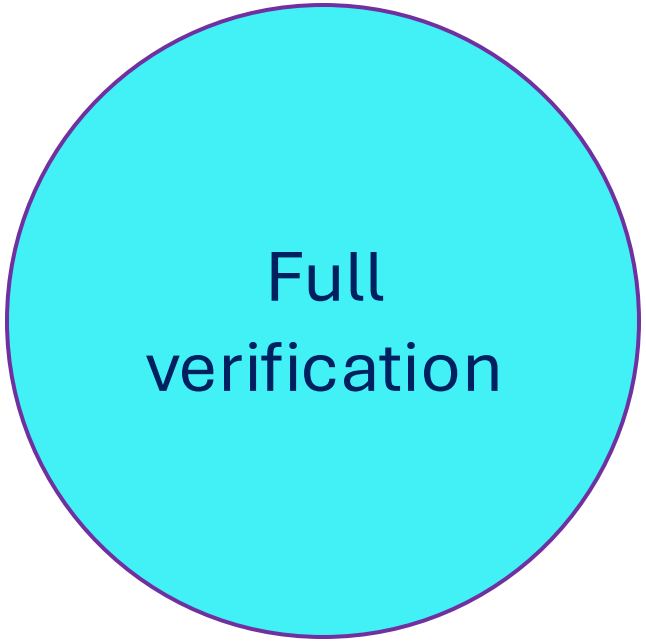
Model M works in the cases that we've tried

Just right?



Theory says convergence → S
+
It seems like we've converged

Too hard



Formal proof that model M obeys spec S

AISI

We're far from asymptotic guarantees currently!

- **Theory expects debate will not converge to approximate honesty**
  - (Same for other existing scalable oversight methods.)

- **Obfuscated arguments** (Barnes 2020)
  - Honesty can be intractably harder than deception in some debate games
  - Worse than "no proof": we **actively expect** scalable oversight to break down as a result

- **Exploration hacking** (Hubinger 2023)
  - Need some reason to believe that converged training implies game equilibria

- **Low vs. high stakes / deceptive alignment** (Christiano 2021)
  - Learning theory will give you at most a 1 – ε success rate. Can we to get to o(1) failures?
  - Sometimes 1 – ε might be enough (sufficiently sandboxed alignment research)

- **There are more problems!** Systematic human errors, etc.

Problem:

Existing scalable oversight methods fail even ignoring deceptive alignment

Hopeful claim:

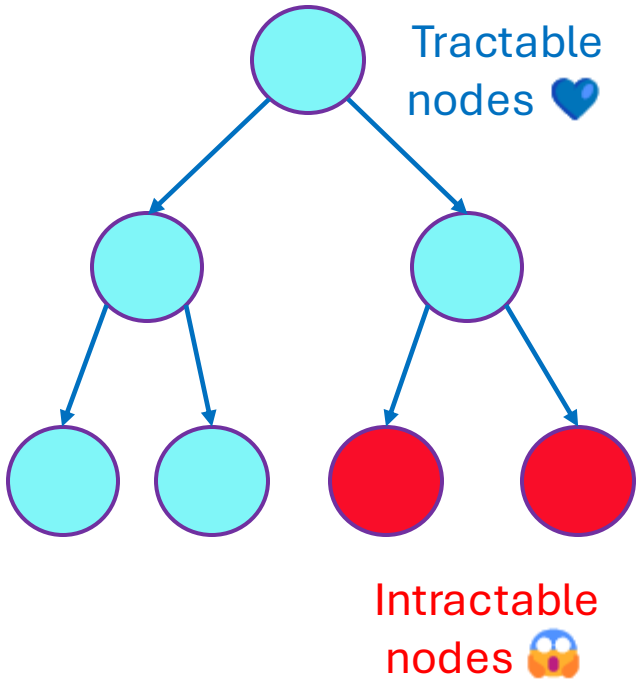
We have a meaningful shot at better theory (including translation to empirics)

Complexity theory and obfuscated arguments

- **Goal:** Use ML to accelerate an expensive computation
  - Write down a huge tree computation
  - Use heuristics to skip a bunch of steps
  - AlphaGo uses this for go, debate uses it for safety

- **Problem:** Obfuscated arguments (Barnes 2020)
  - The heuristics will not always work!
  - **Risk: We drop from a tractable node to intractable nodes**
  - Applies to all scalable oversight methods
    - Amplification, Scientist AI, etc.

- Obfuscation may **be an attractor**
  - Arose naturally in human experiments (Barnes 2020)
  - Winning strategy in easy theory examples



Example: "This kind of Python code likely has vulnerabilities, but we don't know where yet."

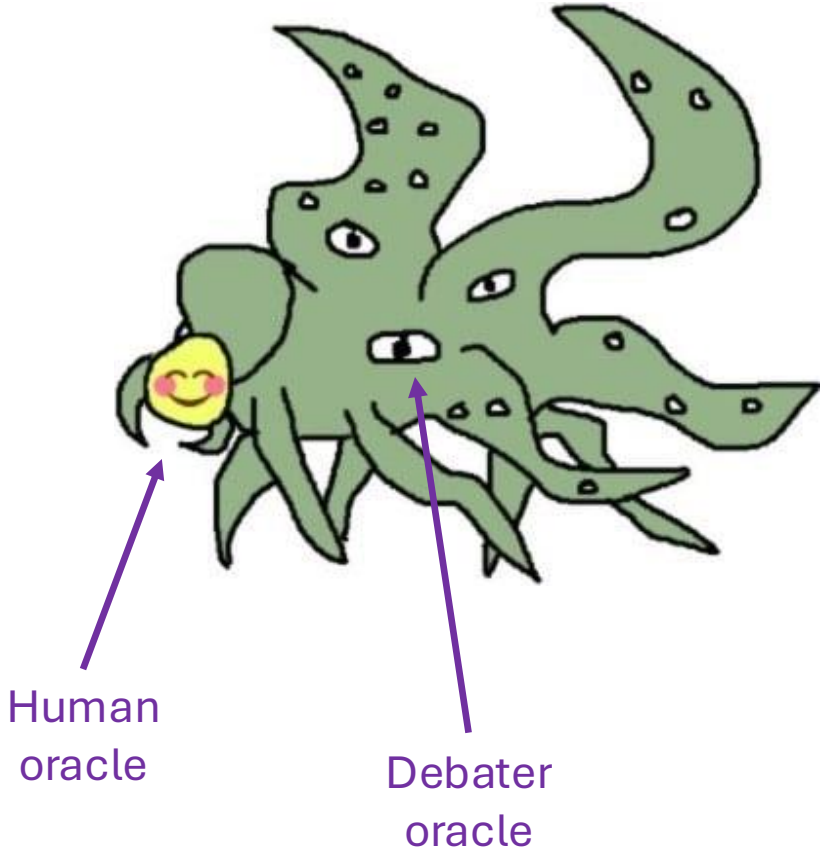
Needs to work despite **fragile, alien heuristics**

- **Subtlety: Tractable is relative to the ML model**
  - ML models can be narrowly superhuman, or much weaker
  - We don't know where in advance!

- **Good news!** Problem is…

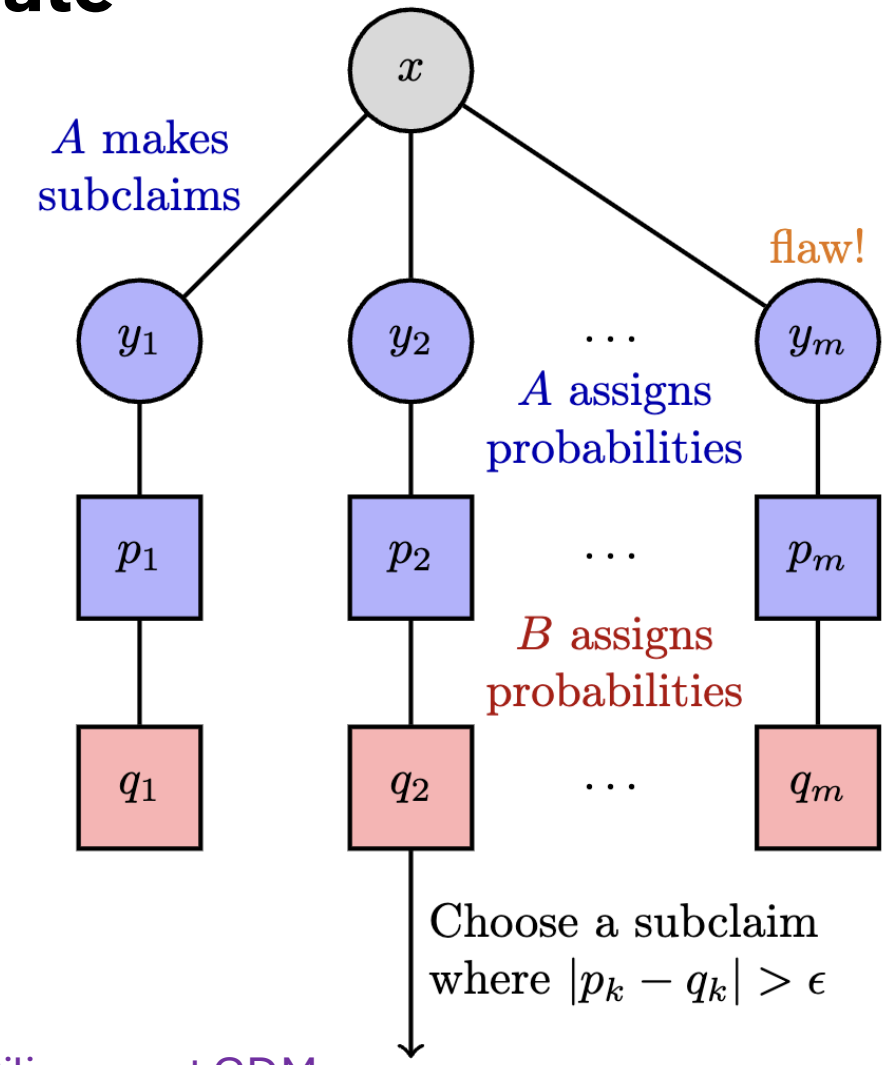
- **Easy to model**, with two oracles (Brown-Cohen, Irving, 2024)
  - **Human oracle** to capture human judge
  - **Debater oracle** to capture alien machine reasoning
  - Protocol can't call the debater oracle directly

- **Novel:** complexity theory hasn't tackled it yet
  - So no reason to believe it's hard!
  - Complexity theory is often either impossible or easy



In-progress attempt: **prover-estimator debate**

- **In original debate:**
  - Alice decomposes the problem
  - Bob chooses where to recurse
- **Bad:** Alice wins if **both** Alice + Bob are confused
- **Prover-estimator debate:**
  - Alice decomposes
  - Alice and Bob assign probabilities
  - If they disagree a lot, recurse
  - If they're close, **use Bob's probabilities**
- **Goal:** Bob wins if they're both confused



Joint work with Jonah Brown-Cohen + Georgios Piliouras at GDM

Two caveats: proof holes + stability

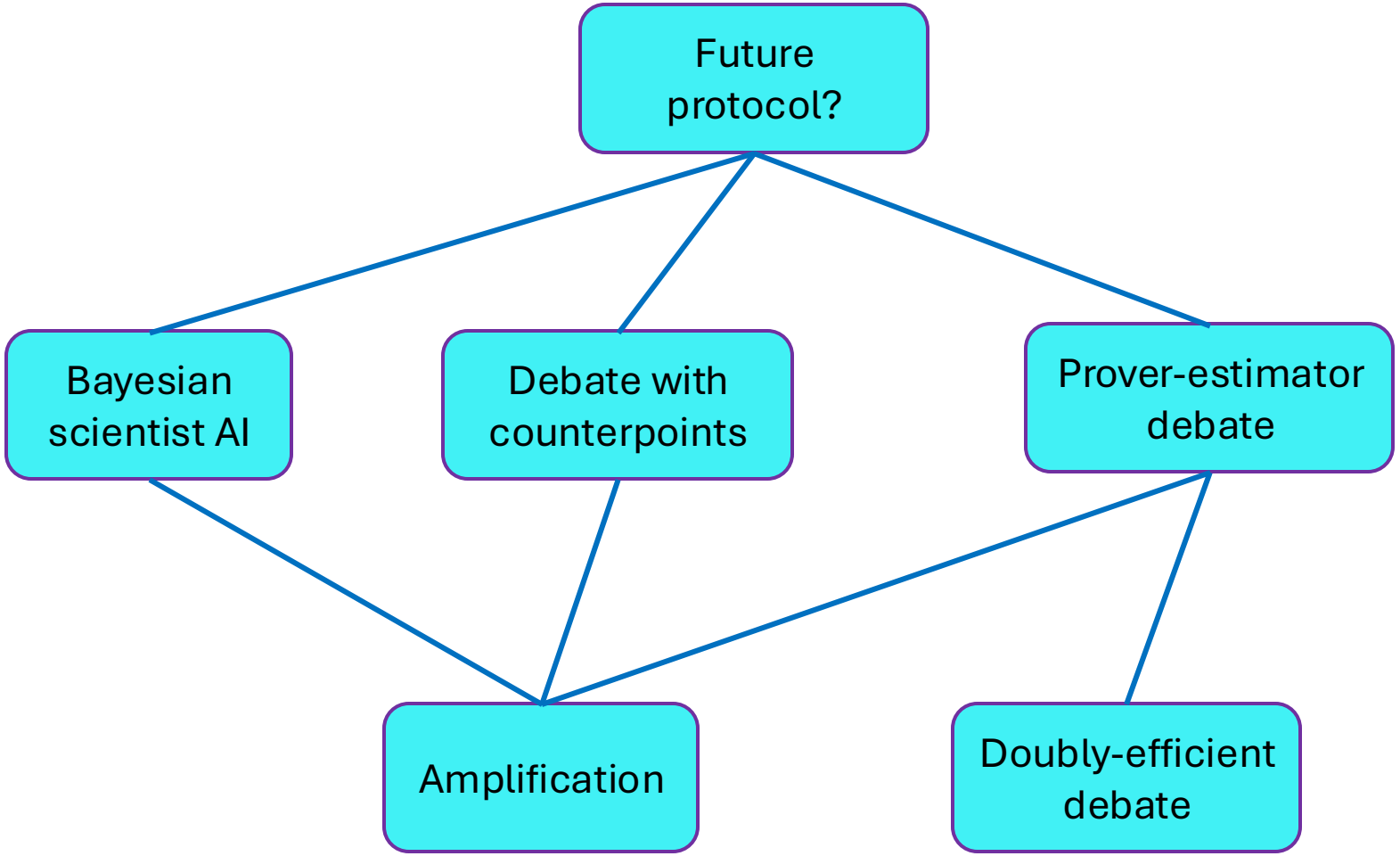
Still nailing down the proof!

- Only a few pages, but super fiddly
- I'm at **80%** on true in ~current form
- Theorem will have messy details
  - $O(\rho^{2d}/(1-\rho)^2 m^2 d^2/\varepsilon^2)$ and such
- But still would be progress!
  - Deals nicely with alien heuristics

Need to assume stability

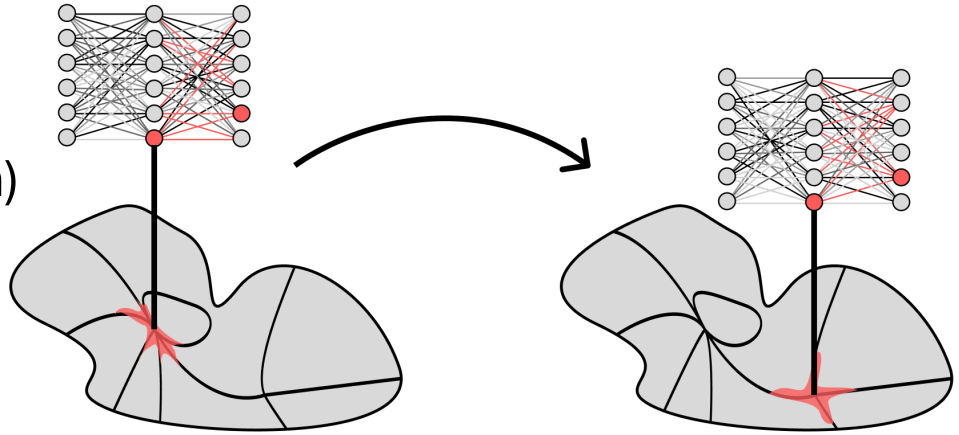
- Small L^∞-changes to child probabilities change node probabilities only a bit
  - Likely required, in some form
- Jonah's optimistic intuition
  - Holds if independent evidence can be found for subnodes
- Beth's pessimistic take
  - Just says when obfuscation applies :)
- **Needed only for capability, not safety**
  - Bob defeats unstable lies

We have many scalable oversight schemes, with different advantages



Learning theory and exploration hacking

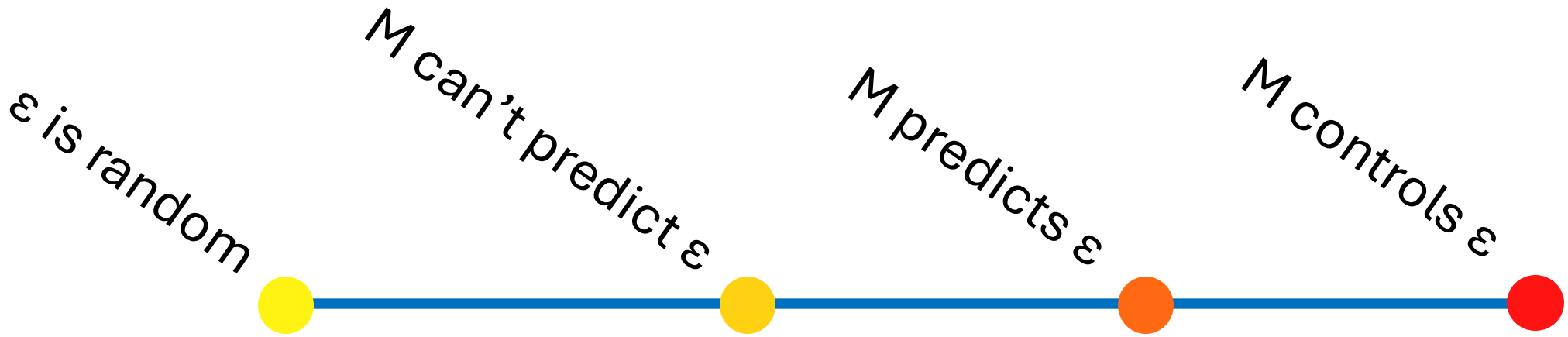
- Learning theory says when we can converge to near-optimum
  - **…if everything is convex**
- Multiple hopes for theory closer to neural nets…
- **Singular learning theory** (Hoogland et al. 2023)
  - Might get us nonlinear Bayesian models (near equilibrium)
  - Substitutes for mech interp by modeling variation in behavior as a function of variation in dataset



- **Deep learning theory**
  - Might get us deep linear models or similar DNN proxies
- **Goal:** Find training algorithms with fewer exploration-hacked equilibria

Can learning theory say something about **residual error**?

- Learning theory might say we get within **ε** of the optimum
- What does that **ε** behaviour look like, according to model M?
  - Some possibilities are worse than others!



This kind of theory will only say fake things are safe

- **Need empirics to confirm our toy models resemble the truth!**

- **Assumptions theory might make** **- Empirical validation**
  - We're almost converged – Do the learning curves say so?
  - Human data is sufficiently accurate – Is it really?
  - The model is exact Bayesian reasoning – Hope we don't depend on that

Theory

Empirics

Analogy: The Lax-Wendroff theorem

- We're numerically solving a hyperbolic PDE, and
  - Our discretization scheme is **conservative**
  - We **converge** to something

- **Theorem (Lax-Wendroff, 1960):**
  - Then we converge to a (weak) solution

- How do we check these two properties?

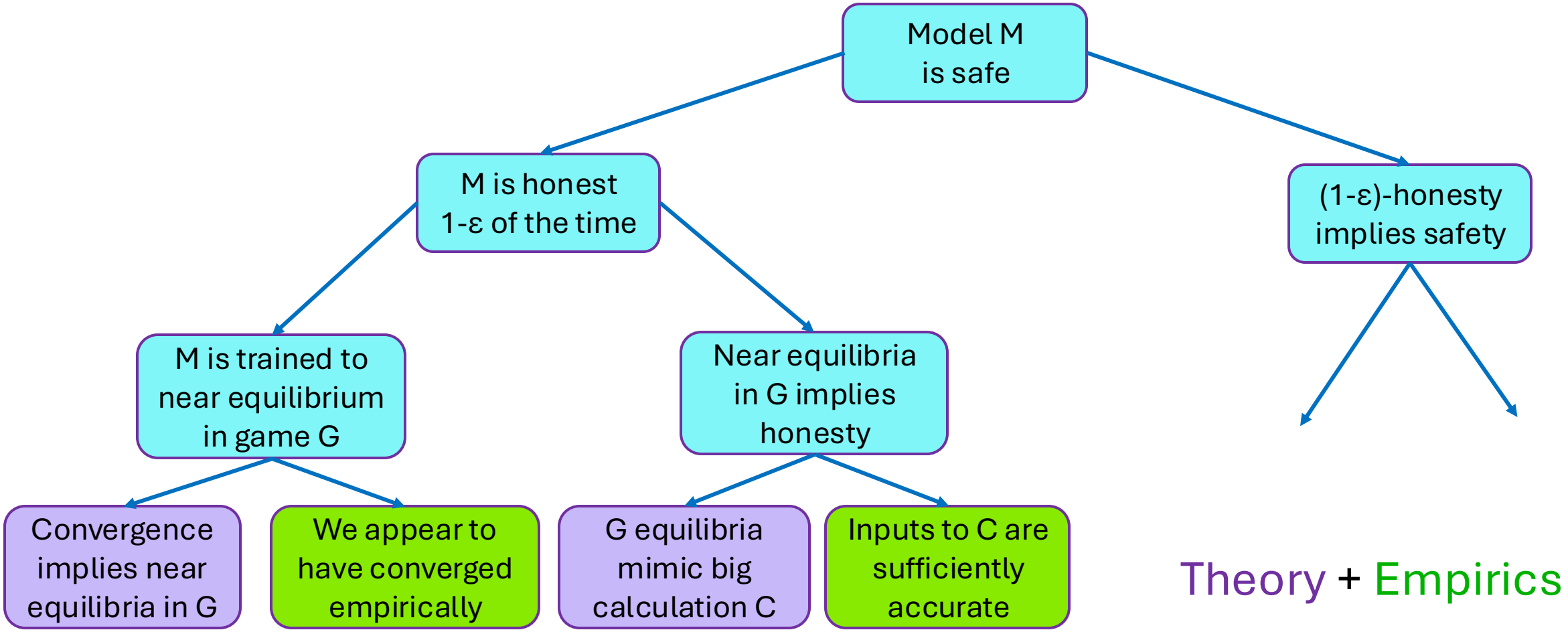
- **Conservative (theory):** Straightforward calculation

- **Convergence (empirics):** Refine the grid a few times. Does it look okay?

Zooming out:

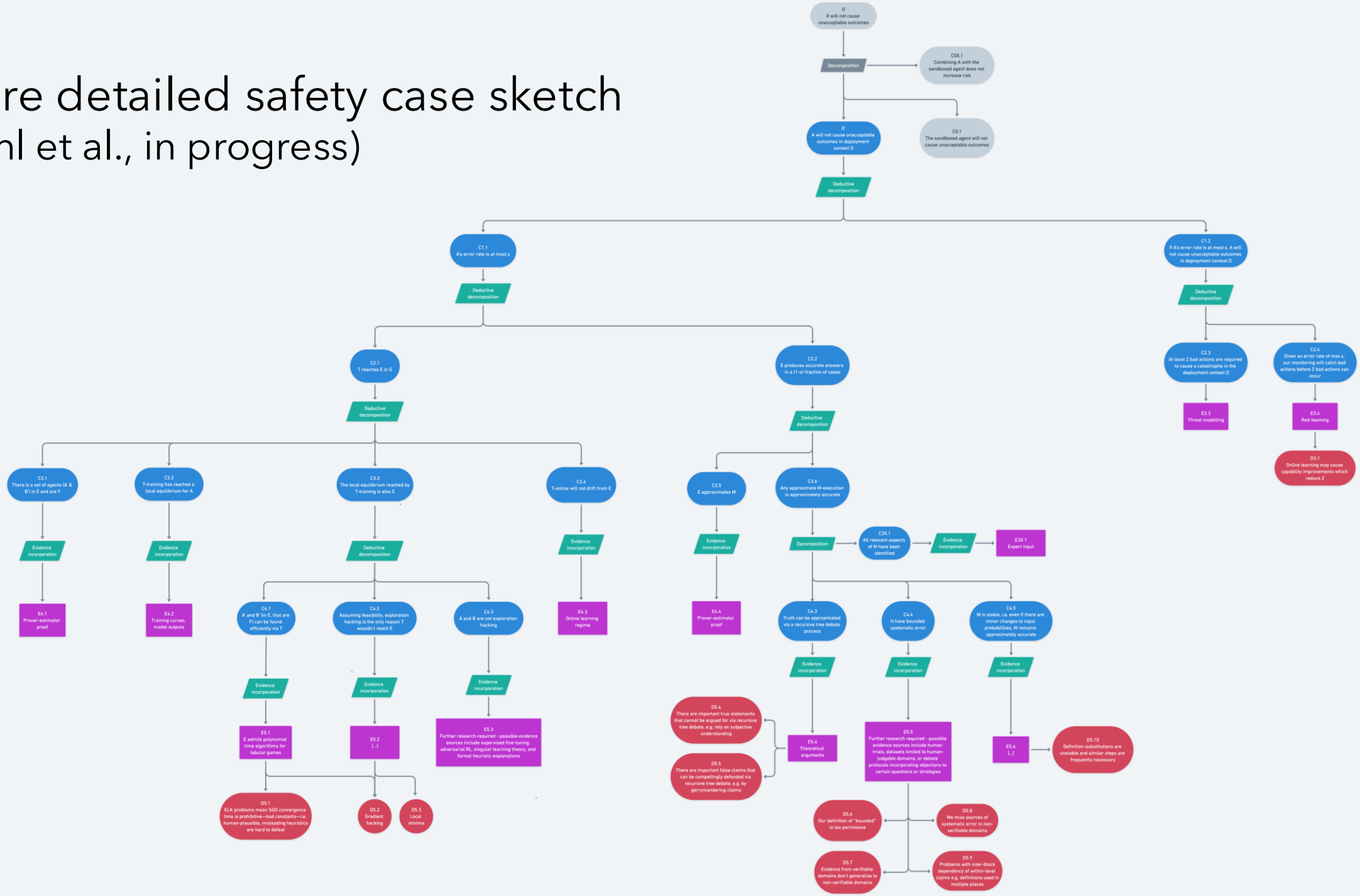
Trying to map all the holes

What asymptotic guarantees could look like for scalable oversight



Theory + Empirics

More detailed safety case sketch
(Buhl et al., in progress)



AISI Alignment Team plan:
Decompose alignment, then fund subproblems in parallel

Sketch safety cases for multiple alignment plans

- First one will be scalable oversight + whitebox
- Will do more internally, and would love to fund external sketches

Fund as many subproblems in parallel as possible

- Tons of people with relevant expertise not yet working on safety
- Independent subproblems → lower entry barriers + parallelism

Fund both theory and empirics

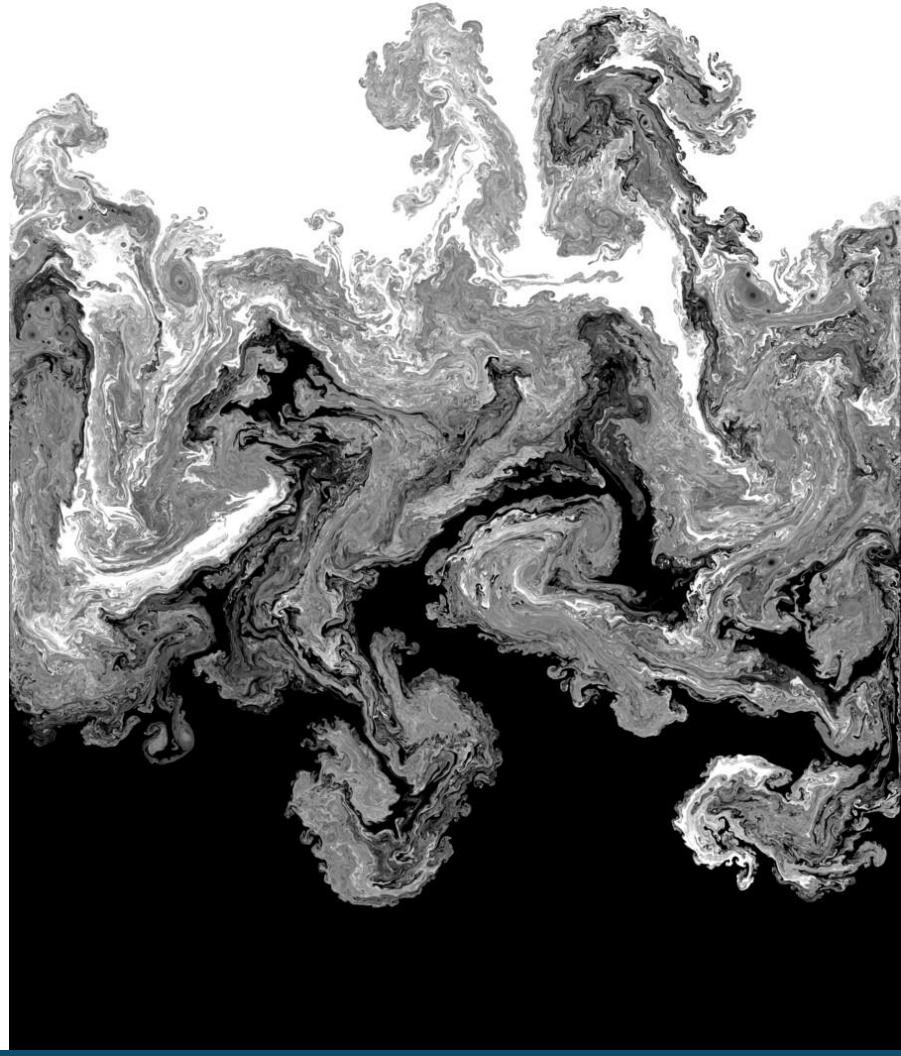
- Theory is neglected + accessible outside labs (academia, nonprofits)
- More theory needs more empirics to target issues theory uncovers

We have funding for alignment research, including theory

https://forms.office.com/e/BFbeUeWYQ9

Mapping the blackbox / whitebox boundary

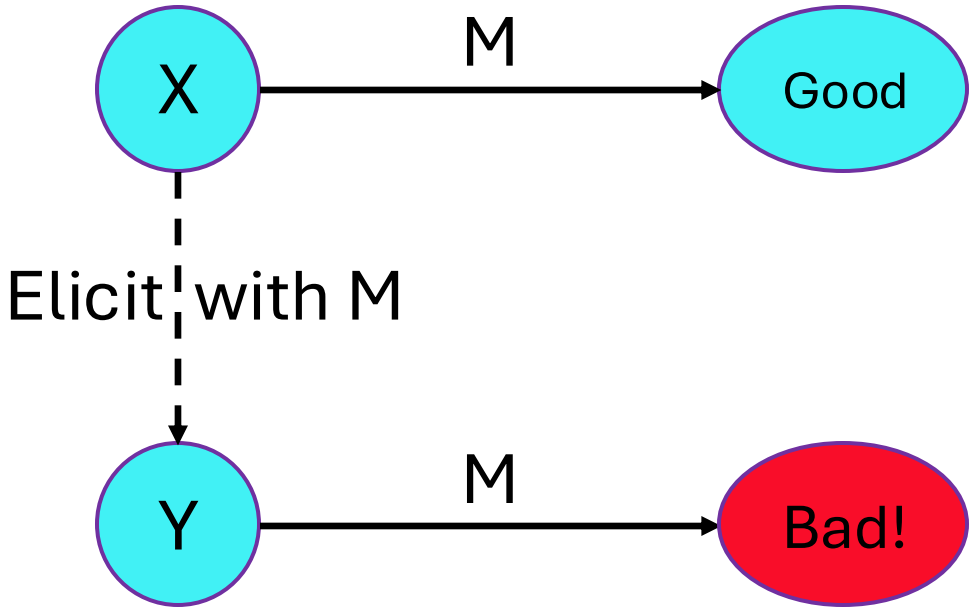
- Many possible levels of interpretability success
  - Linear probes → … → full circuit breakdown
- How far does partial success get us?
- **First:** Describe the lowest level where debate works
- **Second:** Find many independent routes
  - Mech interp, SLT, computational mechanics, self-other overlap, adversarial ML, automated interpretability, …



A candidate level: **eliciting bad contexts**

- On context X, model M is innocuous
- But M “knows” that on context Y, it would defect
- **Can we extract Y?**
- Clearly requires whitebox search, but maybe not full understanding
- Alas, harder than automation might easily provide (Pfau 2025)

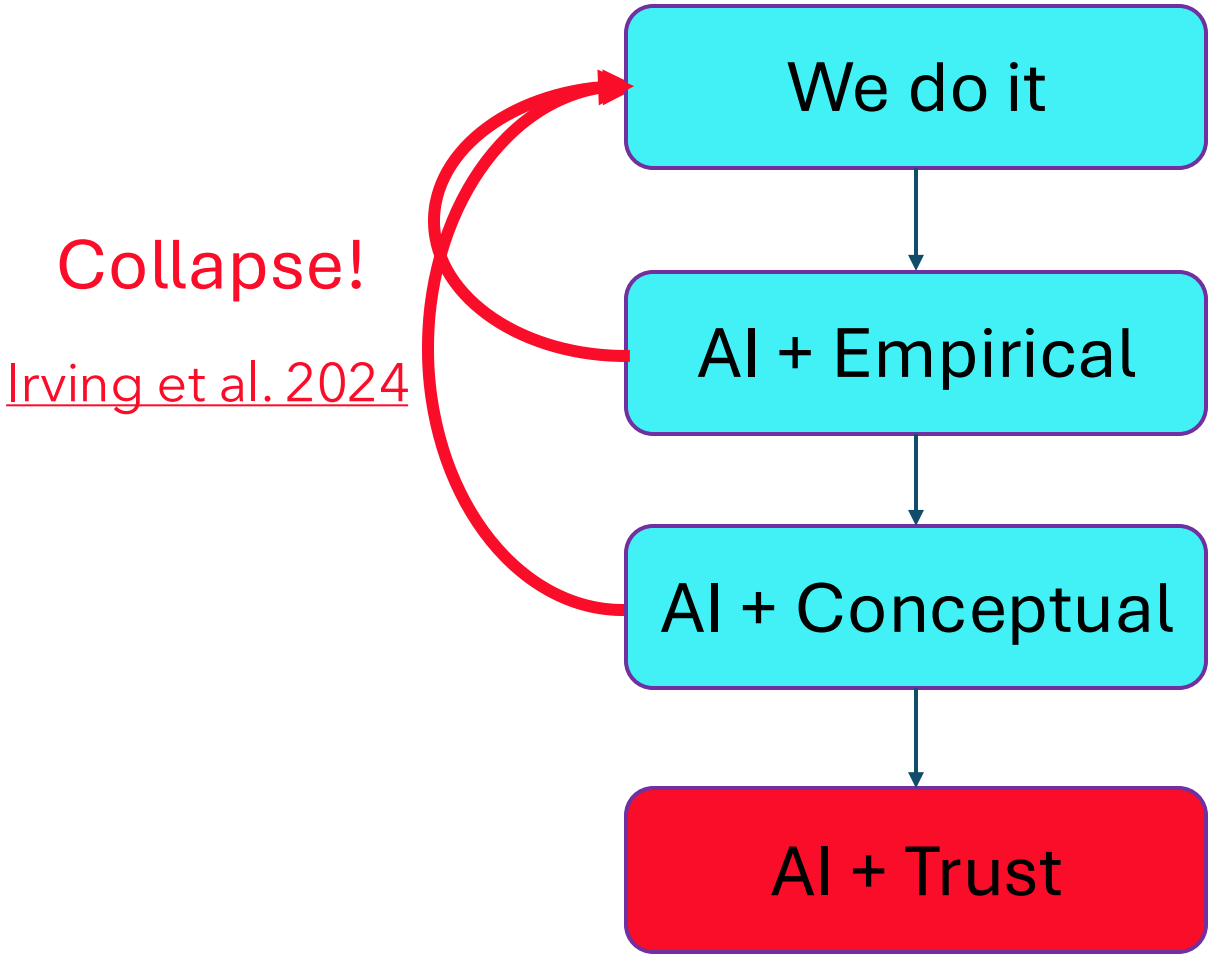
Irving et al. 2025



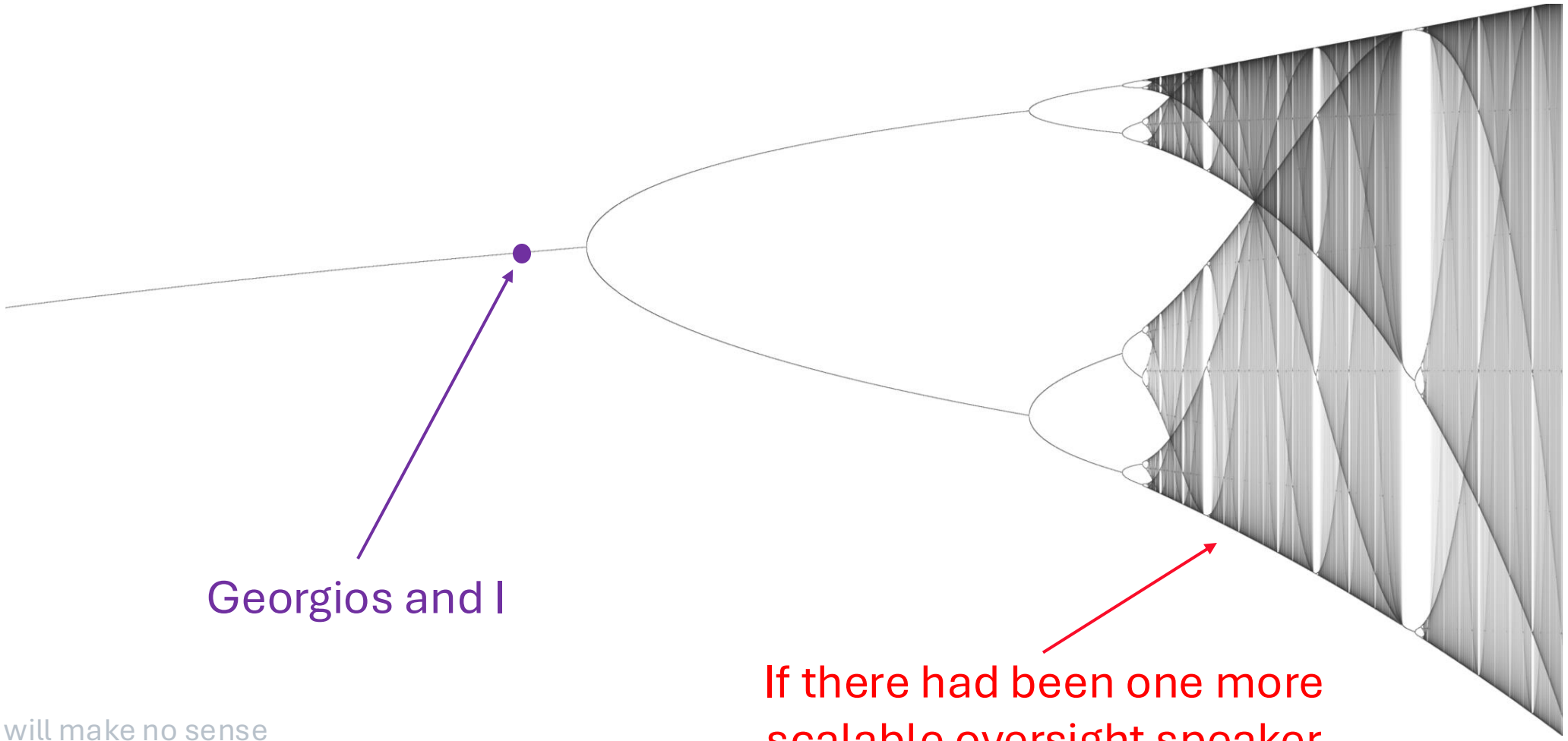
Why not outsource
alignment to the machines?

We might need to understand alignment to outsource it

- Ways to check AI work:
- **Empirical** means we partly know an alignment scheme
- **Conceptual** means we understand, well enough
- **Trust** is bad: AIs make mistakes too!
- **Shouldn't give up trying to solve it ourselves!**



Georgios and I managed to coordinate without communication!



Apologies, slide will make no sense outside of the Simons workshop...

Thank you!

AISI alignment research funding: