Perspectives: Opportunities at the Intersection of FM & AI

Nora Ammann, 2025

$ymposium on AI ✓erification

Before we start...

- I'm not an AI Verification expert!
- Background in AI safety & security
- UK's Advanced Research & Invention Agency
    - **Maths for Safe AI**
    - **Safeguarded AI**

Mathematical proof is the gold standard of confidence and assurance. **How much can we use these tools to make AI safe?**

Safeguarded AI programme aims to develop a workflow for leveraging general-purpose AI to produce domain-specific AI applications with **quantitative guarantees of safety** in their contexts of use.

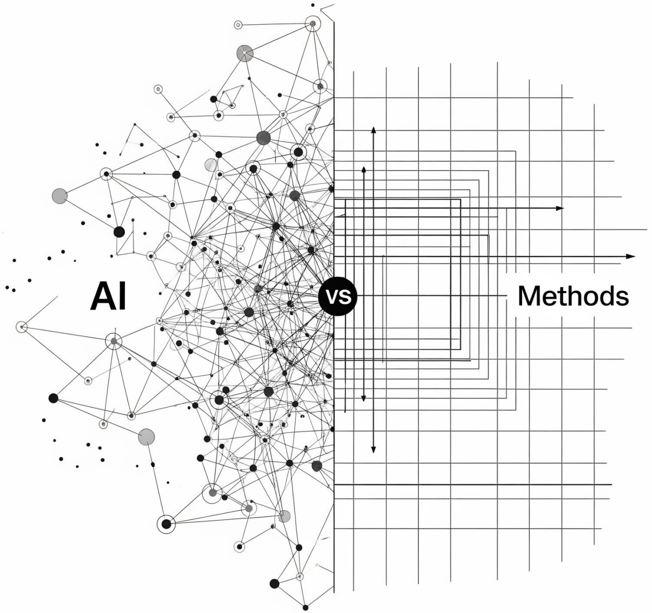
Outline

1. **"Old Rivals, New Friends"** – Rethinking the Synergies of FM & AI
2. **A Space of Opportunities** – Secure Software, Safe AI
3. **Call to Action** – Looking for science entrepreneurs!

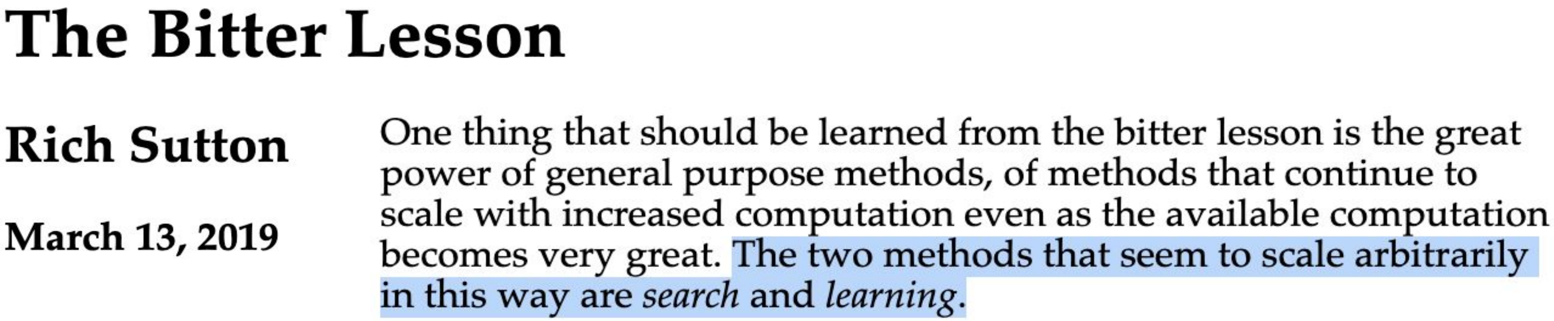
Old Rivals, New Friends

Residual Misgivings

- AI
  - Unprincipled
  - Unreliable, untrustworthy
  - No natural allie in the search of certainty
- FM
  - Do not scale
  - Do not generalise easily / brittle
  - Impractical, hard to use
- Current FM*AI applications
  - Limited use (e.g. input-output, narrow cases)
  - Challenge to scale to complex real-world deployments



Rethinking the FM & AI Synergies



The Bitter Lesson

Rich Sutton

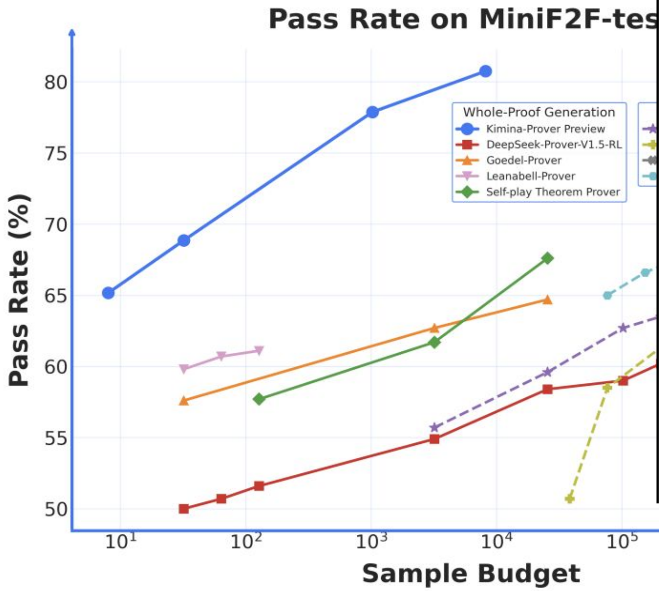
March 13, 2019

One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great. The two methods that seem to scale arbitrarily in this way are *search* and *learning*.

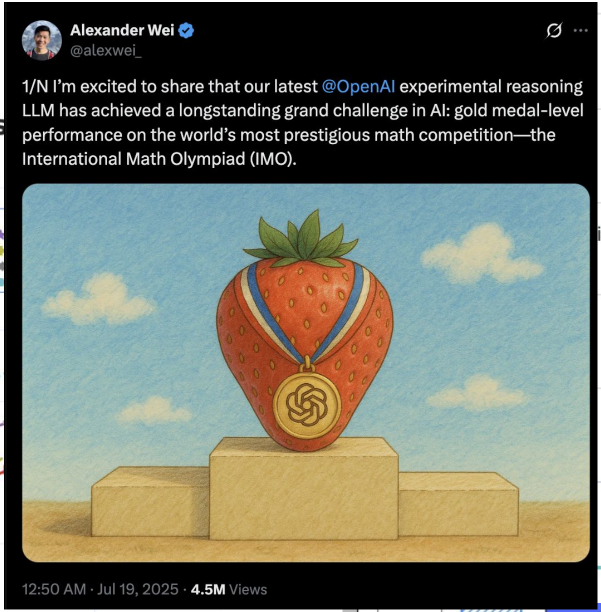
- ➢ Search proofs
- ➢ Search programmes
- ➢ Learn proofs (certificates)
- ➢ Learn translations (informal< >formal, between languages)
- ➢ …

Rethinking the Synergies

- ➢ Search proofs



Wang et al. 2025

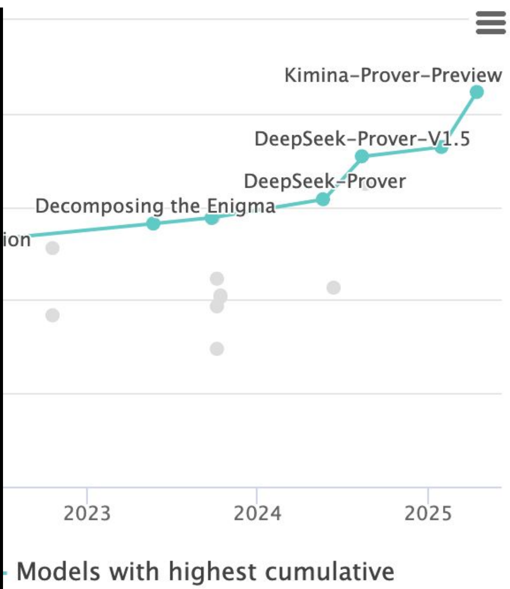


Alexander Wei @alexwei_

1/N I'm excited to share that our latest @OpenAI experimental reasoning LLM has achieved a longstanding grand challenge in AI: gold medal-level performance on the world's most prestigious math competition—the International Math Olympiad (IMO).

12:50 AM · Jul 19, 2025 · **4.5M** Views

Automated Theorem Proving on miniF2F-test



Models with highest cumulative

Rethinking the Synergies

➢ **Learn proofs /proof certificates**

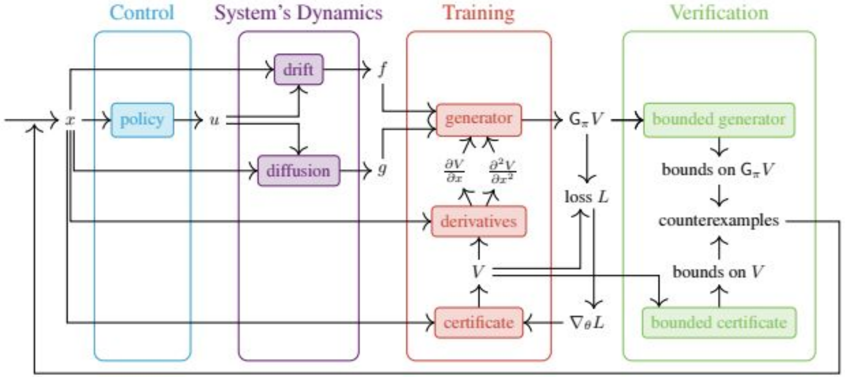
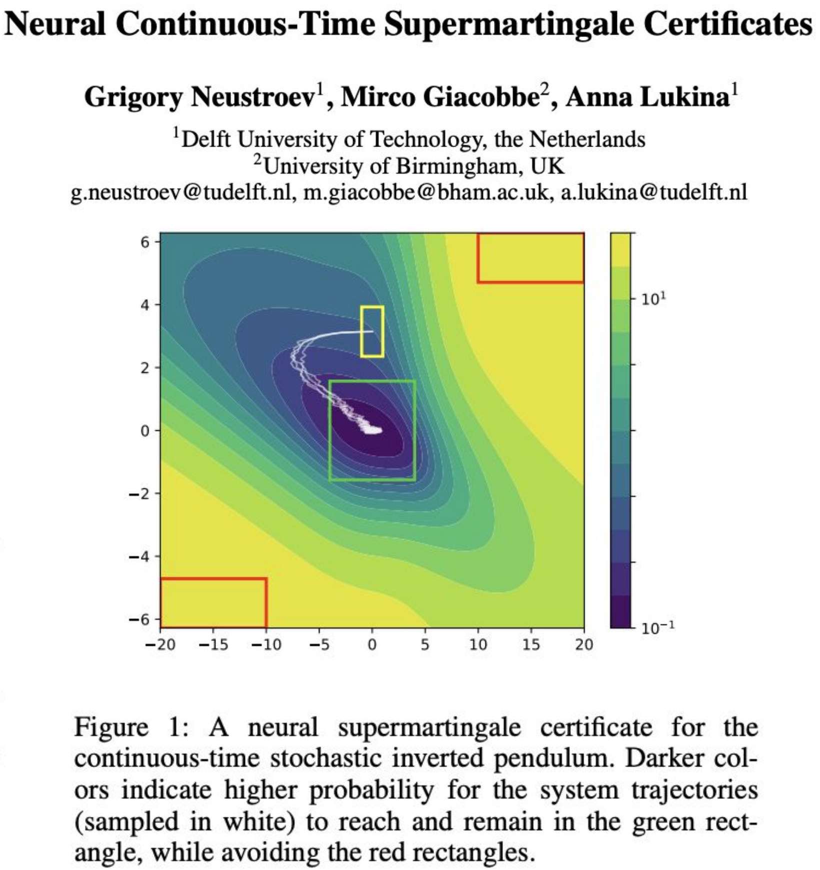


Figure 3: Architecture of the neural certificate training and verification system.



Neural Continuous-Time Supermartingale Certificates

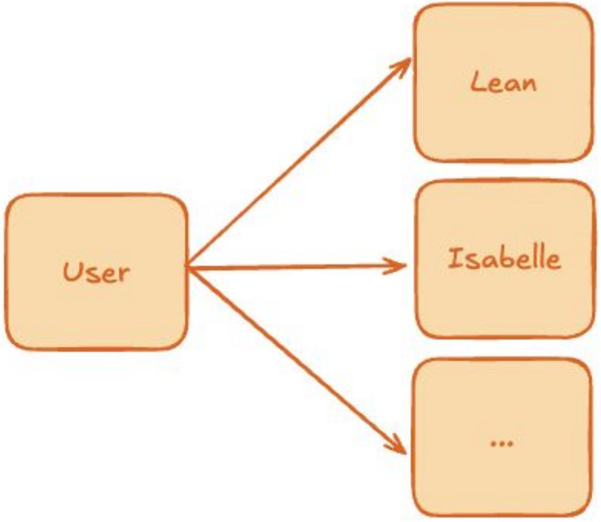
Grigory Neustroev[1], Mirco Giacobbe[2], Anna Lukina[1]

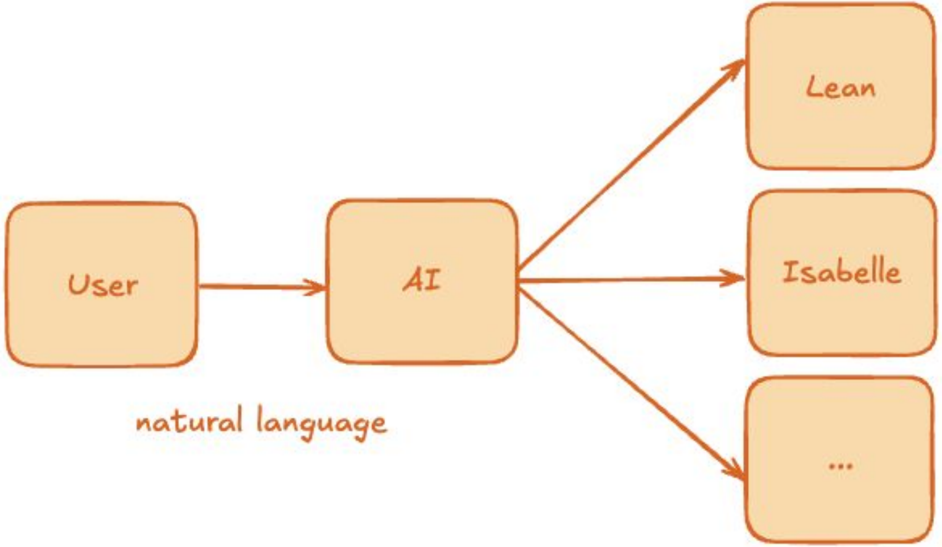
[1]Delft University of Technology, the Netherlands
[2]University of Birmingham, UK
g.neustroev@tudelft.nl, m.giacobbe@bham.ac.uk, a.lukina@tudelft.nl

Figure 1: A neural supermartingale certificate for the continuous-time stochastic inverted pendulum. Darker colors indicate higher probability for the system trajectories (sampled in white) to reach and remain in the green rectangle, while avoiding the red rectangles.

Rethinking the FM*AI Synergies

- ‘Democratize’ formal methods
- Improve adoption





Rethinking the Synergies

AI for FM

- Address proof complexity
- ‘Democratize’ formal methods
- Drive adoption

< >

FM for AI

- Provide rigorous assurance
- Enable responsible adoption

A Space of Opportunities

Secure Software

Challenge

AI is taking software by storm

Coding assistants, ‘vibe coding’, coding agents, etc.

But: AI tends to make code *less* secure (e.g. Chong et al. 2024)

- Introduces bugs
- False sense of security
- Harder to fix

Solutions?

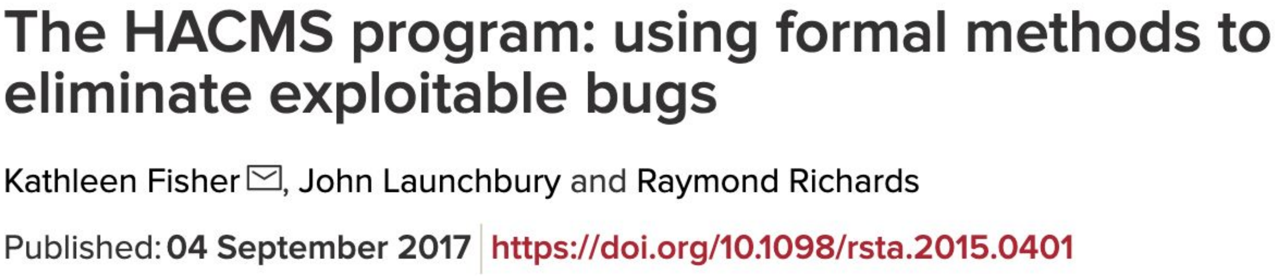
Can we make it *easy*, *cheap* and *the default* to write secure code?

What new affordances does AI give us to do that?

AI-Assisted Formally Verified Code

1. We know how to write secure code.

seL4



The HACMS program: using formal methods to eliminate exploitable bugs

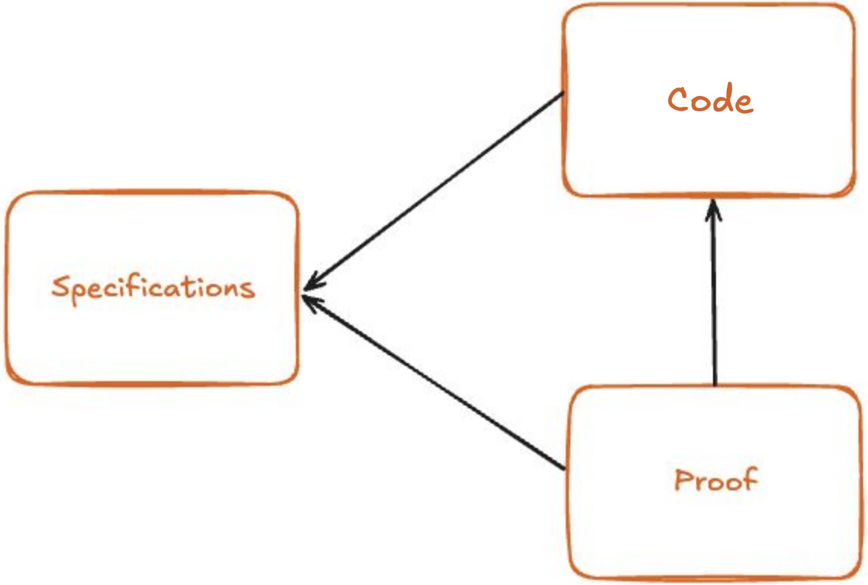
Kathleen Fisher ✉, John Launchbury and Raymond Richards

Published: 04 September 2017 | https://doi.org/10.1098/rsta.2015.0401

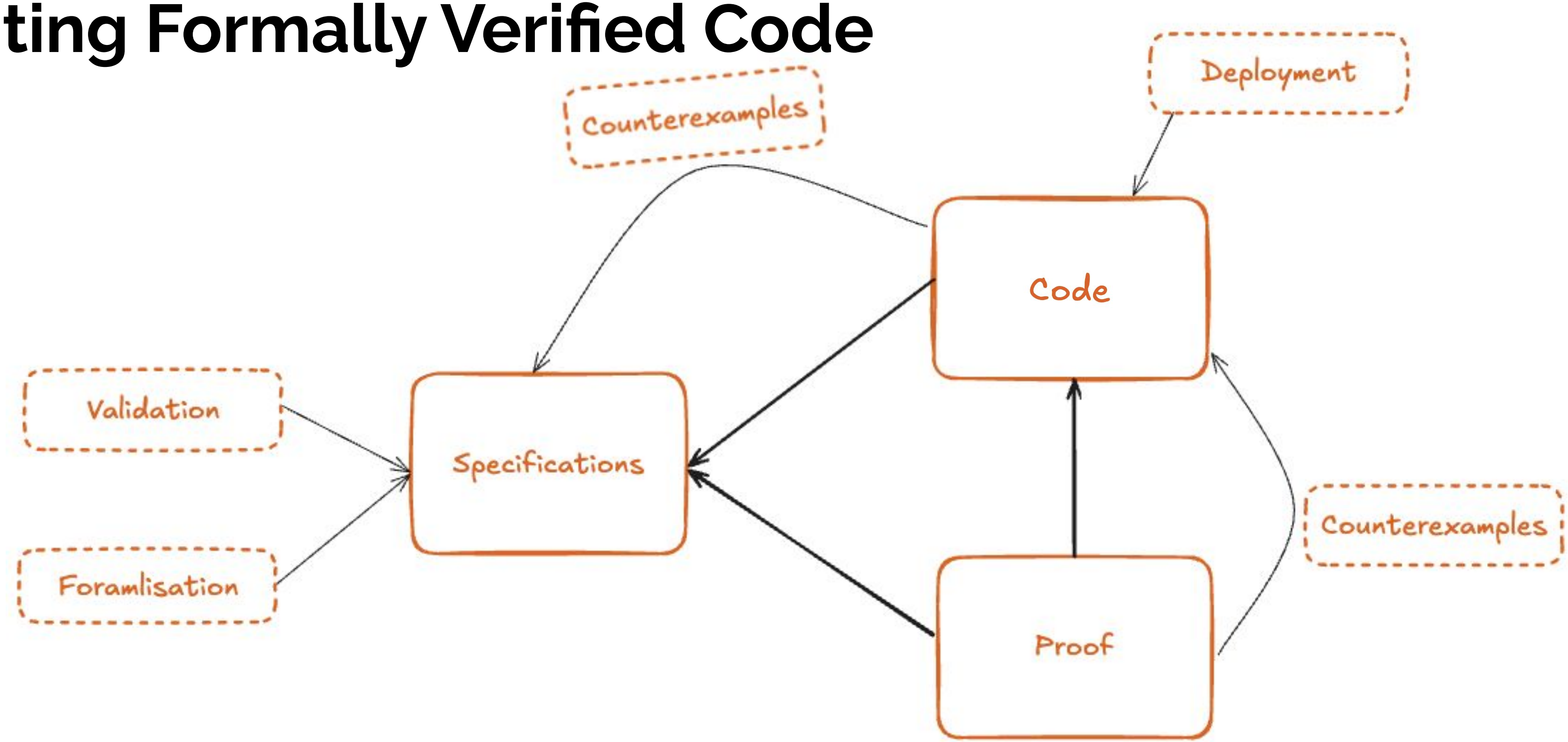
AI-Assisted Formally Verified Code

1. We know how to write secure code.
2. But it takes a lot of human hours!
   a. SeL4 estimated to take 25-30 person years.
3. Let AI do the (hard) work

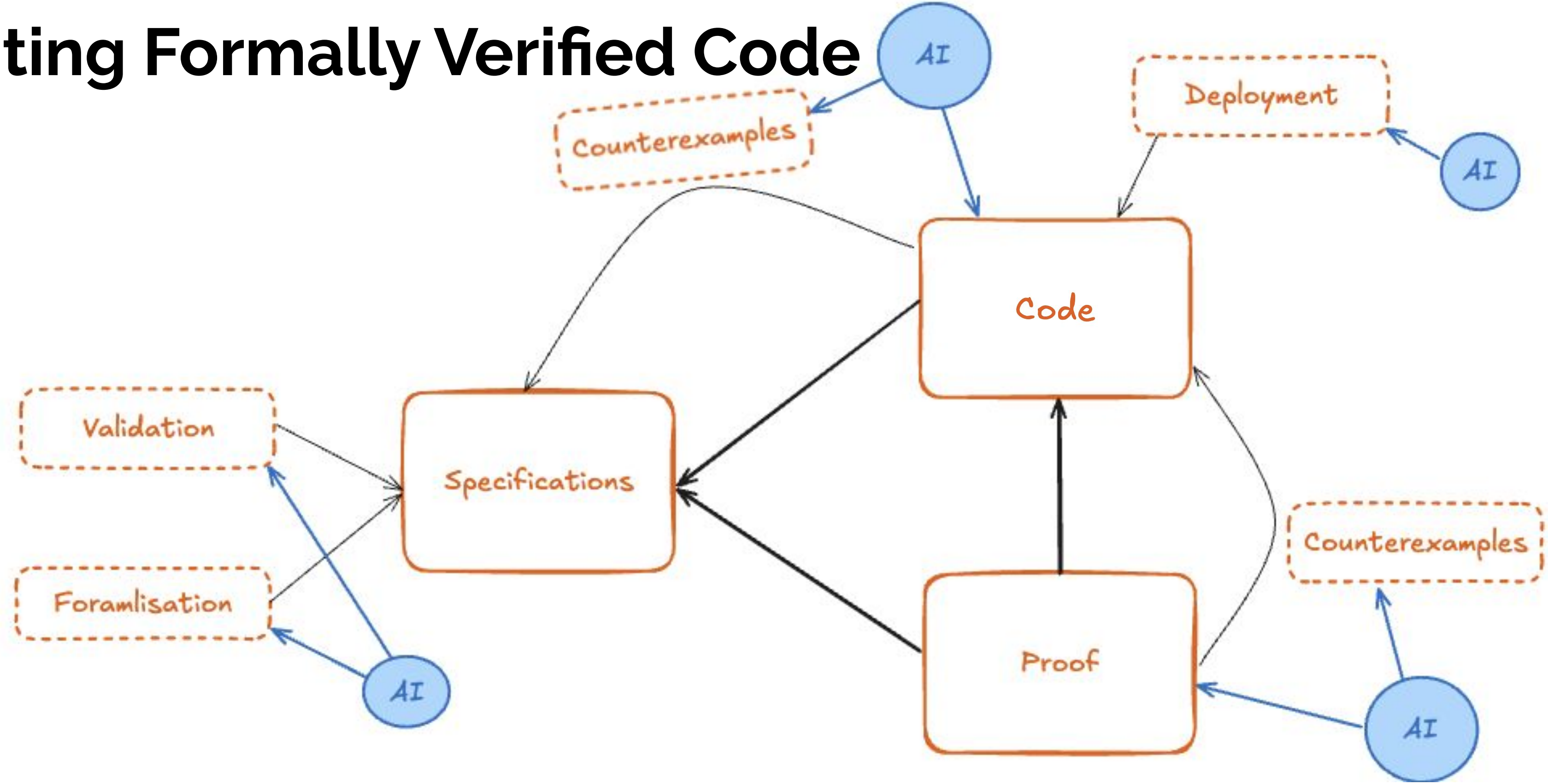
Writing Formally Verified Code



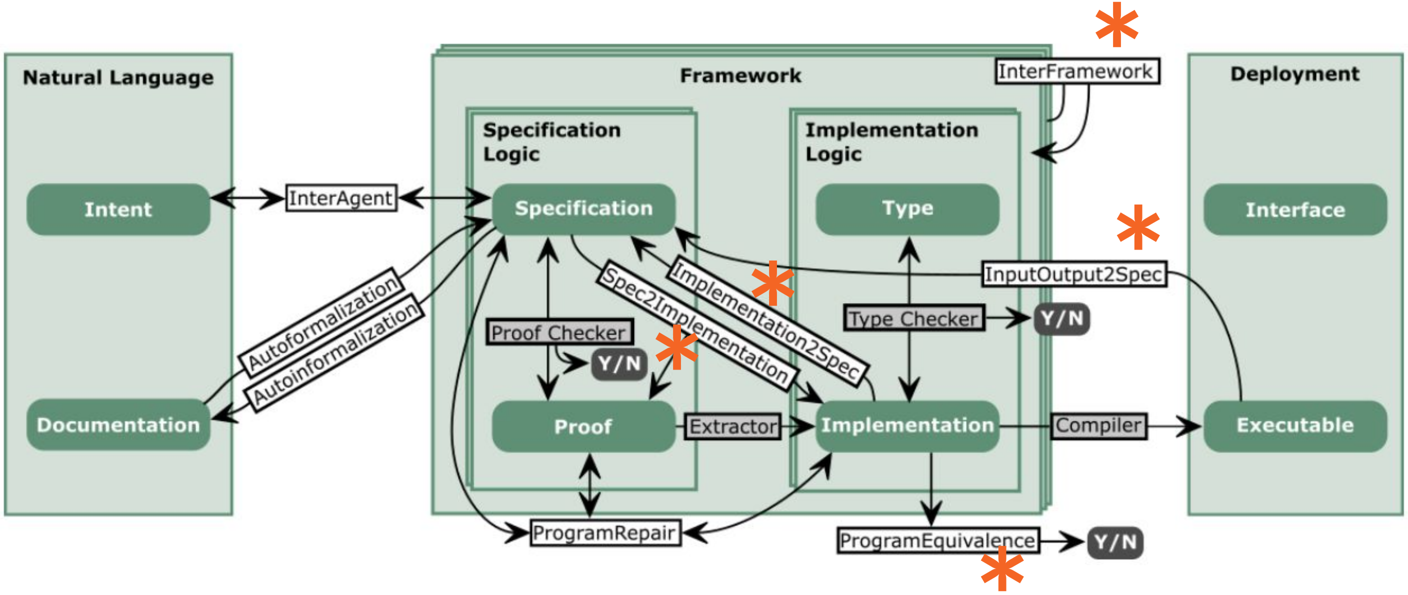
Writing Formally Verified Code



Writing Formally Verified Code



Writing Formally Verified Code



“A Toolchain for AI-Assisted Code Specification, Synthesis and Verification”

Atlas Computing, Lin et al. 2024

Safe AI

Challenge

Rapid progress in AI

Incentive to deploy blackbox systems in increasingly consequential contexts

Critical infrastructure, e.g.

- Energy
- Communication
- Transport
- Digital
- Finance

Autonomous systems (transport, military, …)

Misuse risk (e.g. cyber, bio…)

Safe AI

Challenge

Rapid progress in AI

Incentive to deploy blackbox systems in increasingly consequential contexts

Our ability to adequately assess and secure these deployments remains poor

In particular, the challenge of generality and/or autonomy

Impossibility of exhaustive empirical testing for general domains

Classical safety engineering methods depend on defining a **clear ‘operational envelope’** – which is hard to do for general autonomous AI.

Distributional shift; among others caused by AI adoption itself

Safe AI

Challenge

Rapid progress in AI

Incentive to deploy blackbox systems in increasingly consequential contexts

Our ability to adequately assess and secure these deployments remains poor

In particular, the challenge of generality and/or autonomy

How will this play out?

- World 1: Threshold for acceptable risk stays constant, adoption occurs accordingly
- World 2: 'Capability overhang' (capability outpaces assurance) → Increasing pressure to deploy

⇒ Need to upgrade our safety engineering 'machinery' (fast)

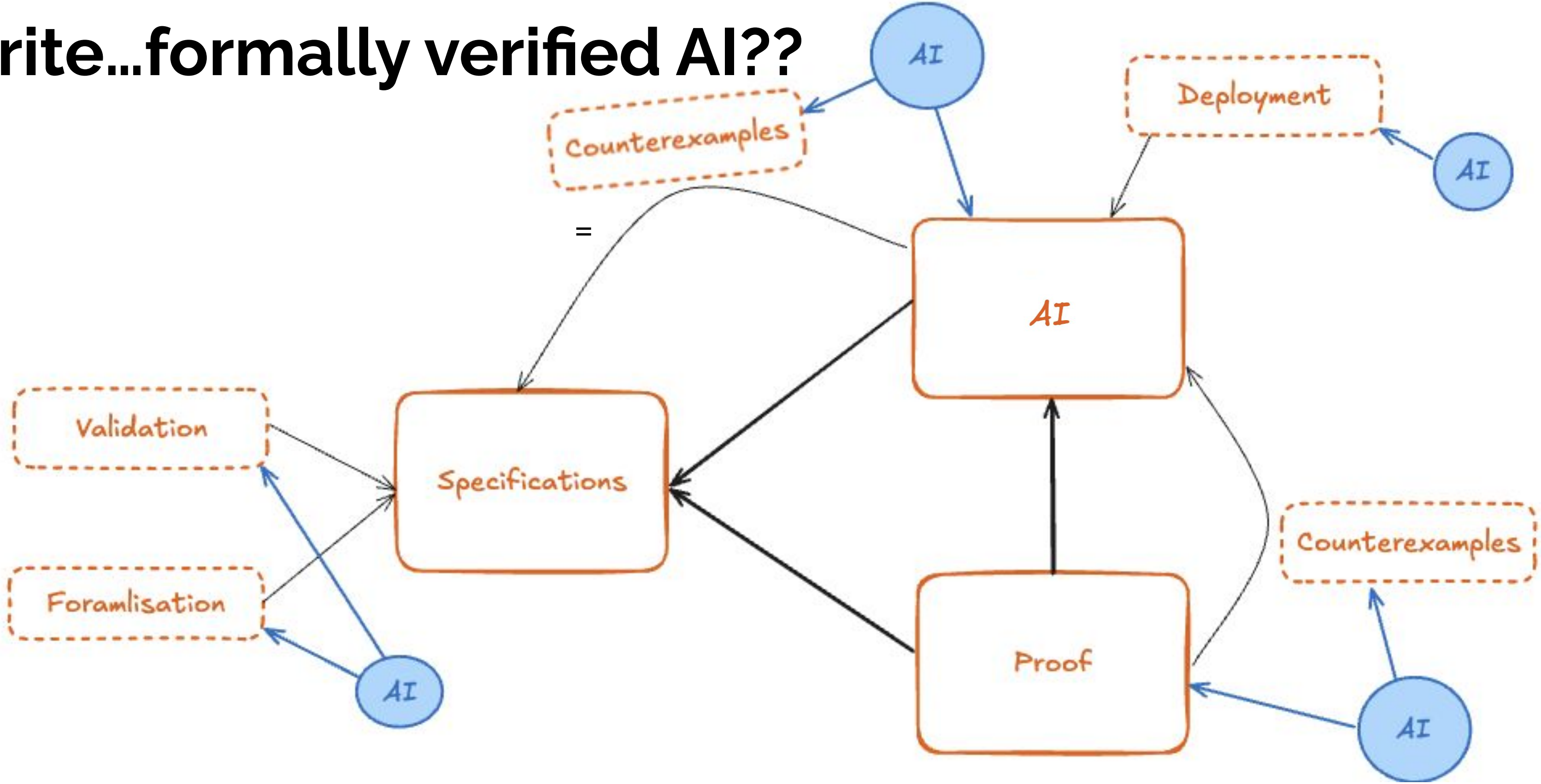
Safe AI

Solutions?

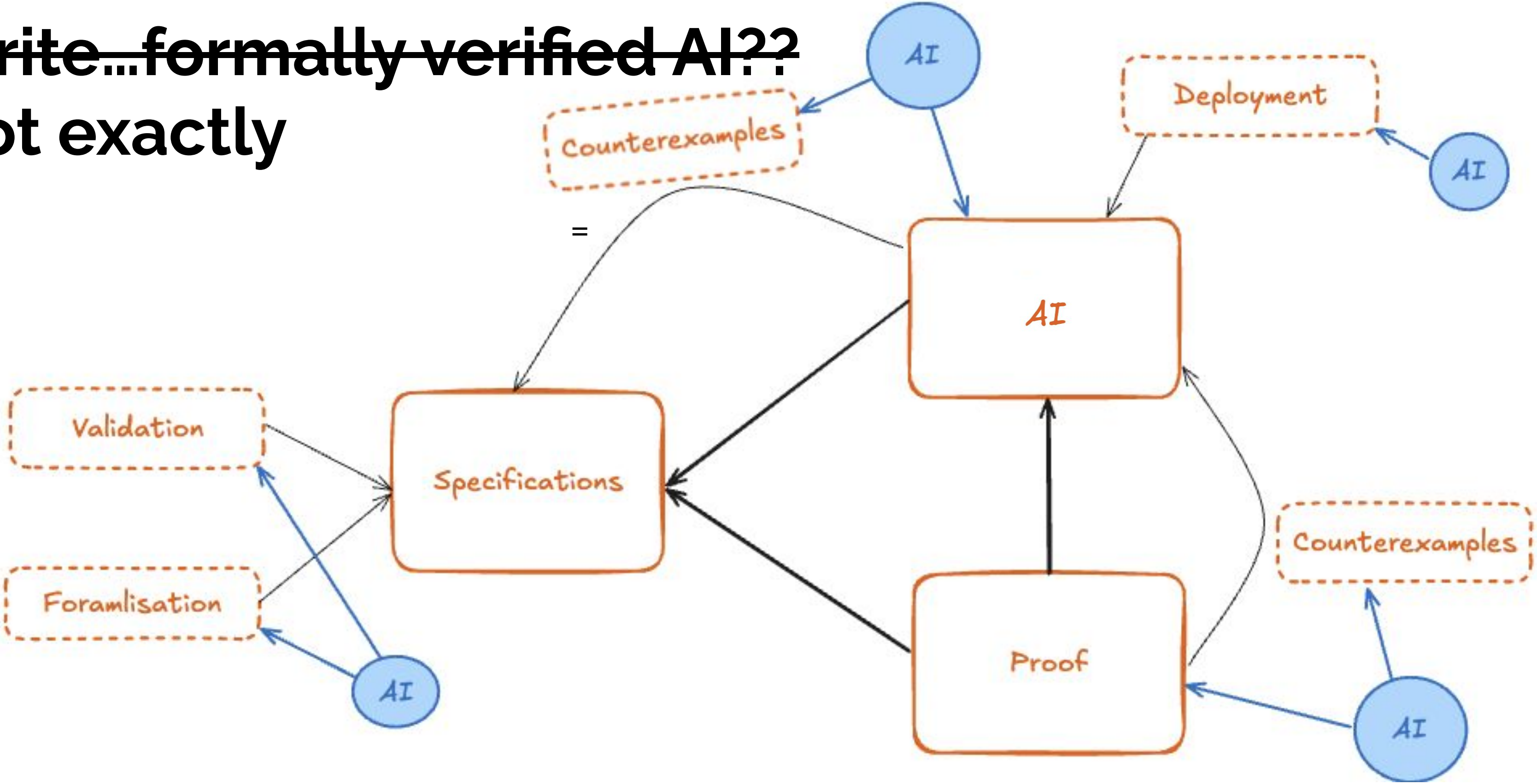
Can we make it *easy*, *cheap* and *the default* to ~~write secure code~~ make safe AI?

What new affordances does AI give us to do that?

Write…formally verified AI??



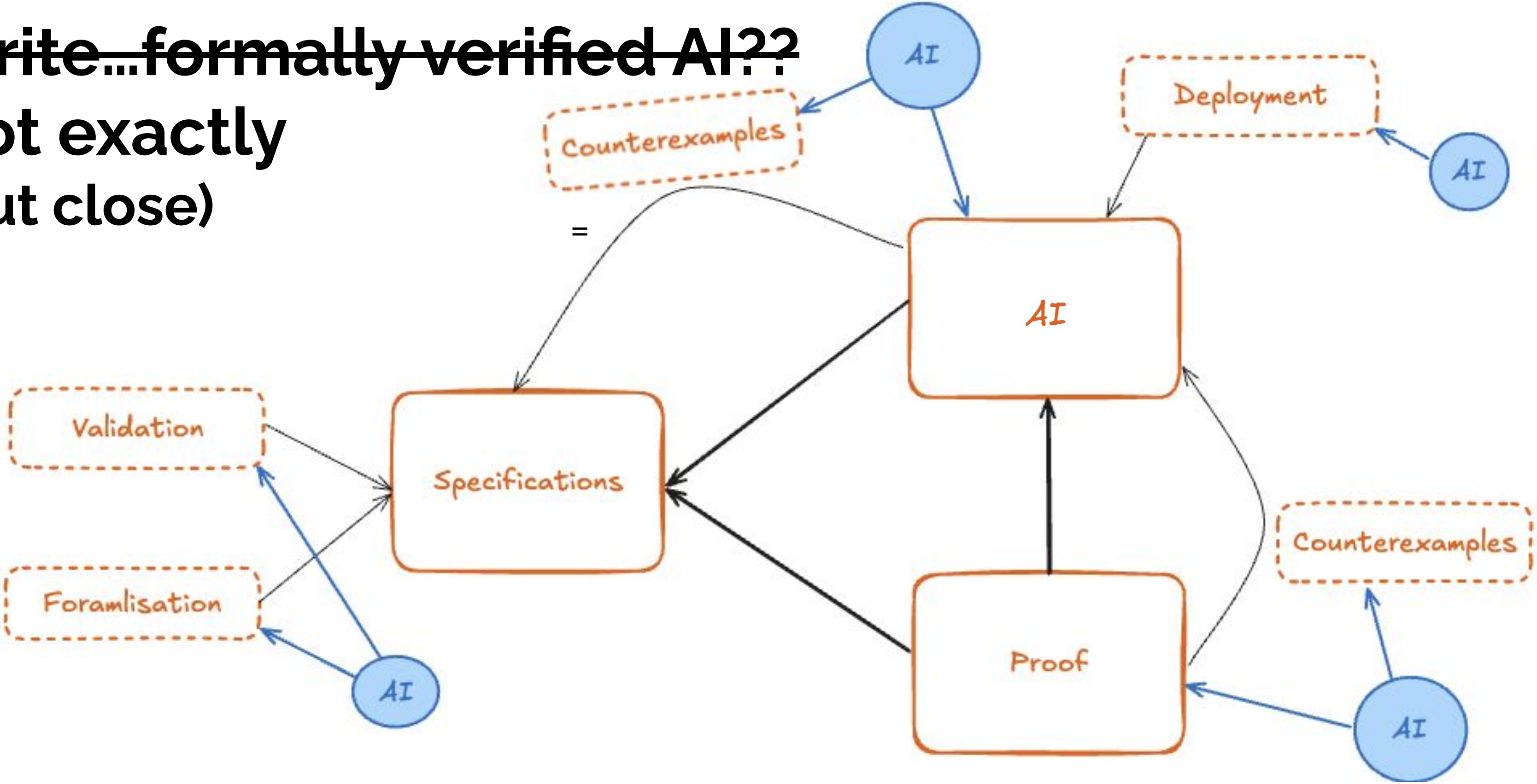
~~Write...formally verified AI??~~ Not exactly



~~Write…formally verified AI??~~

Not exactly

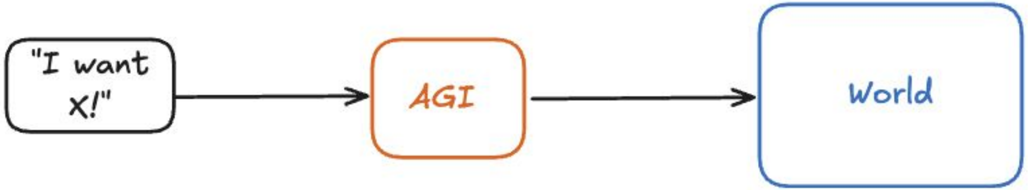
(But close)



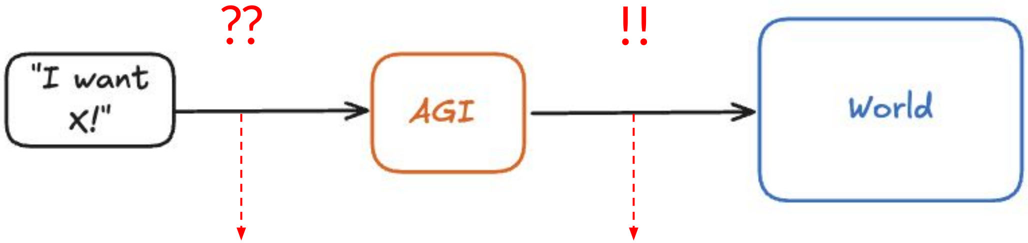
~~Write…formally verified AI??~~
Not exactly
(But close)

Get AI to write a {AI application} that it can certify is correct.

Safe AI



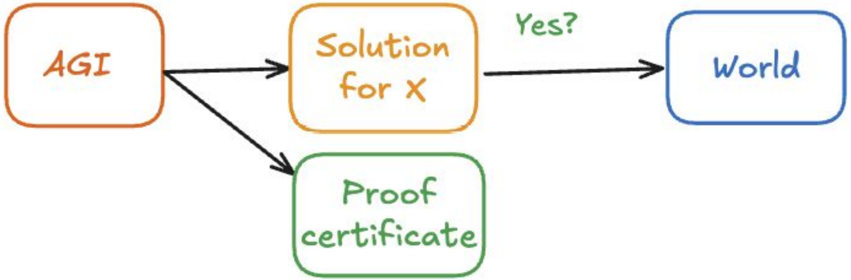
Not-Safe AI



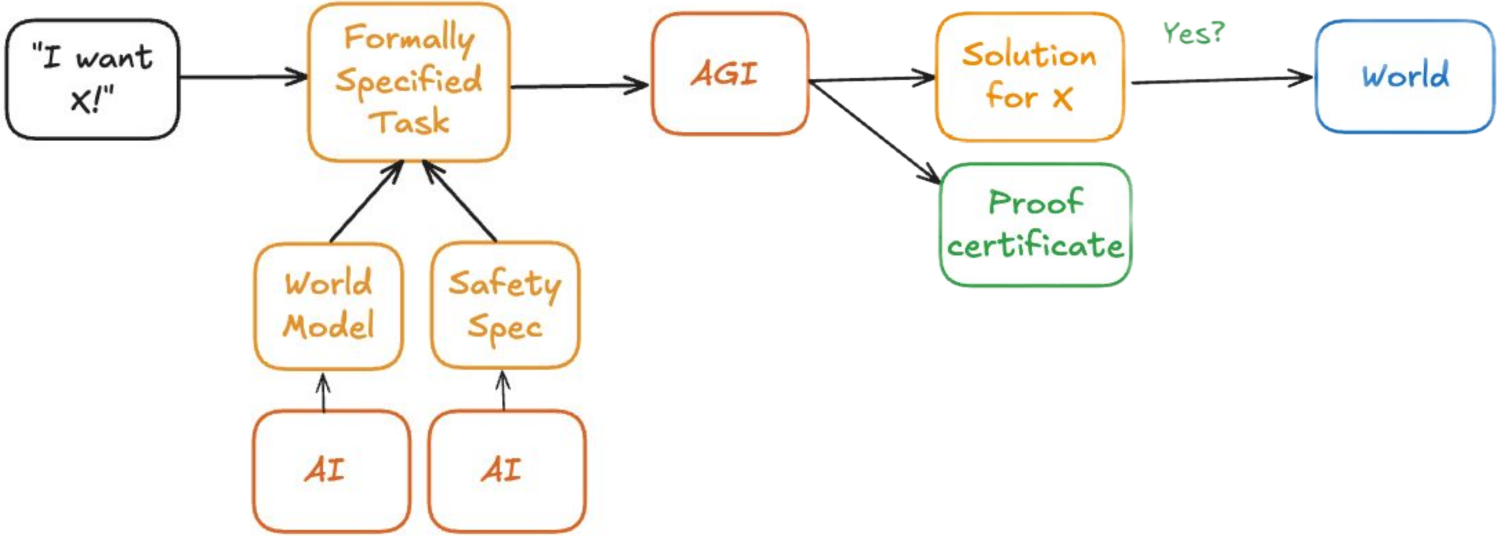
What do we mean?

Distributional shift
Side-effects
Scheming
Misuse
Multi-agent risks
...

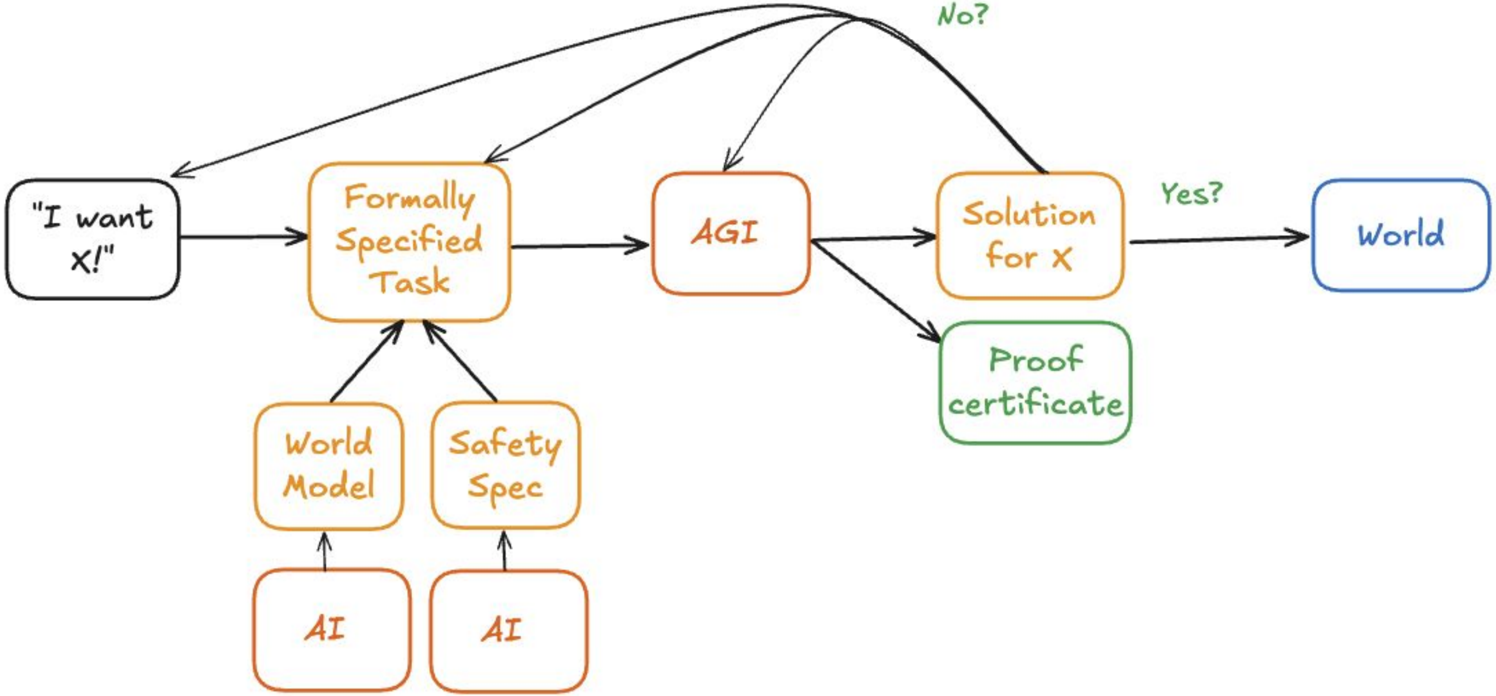
Safe AI?



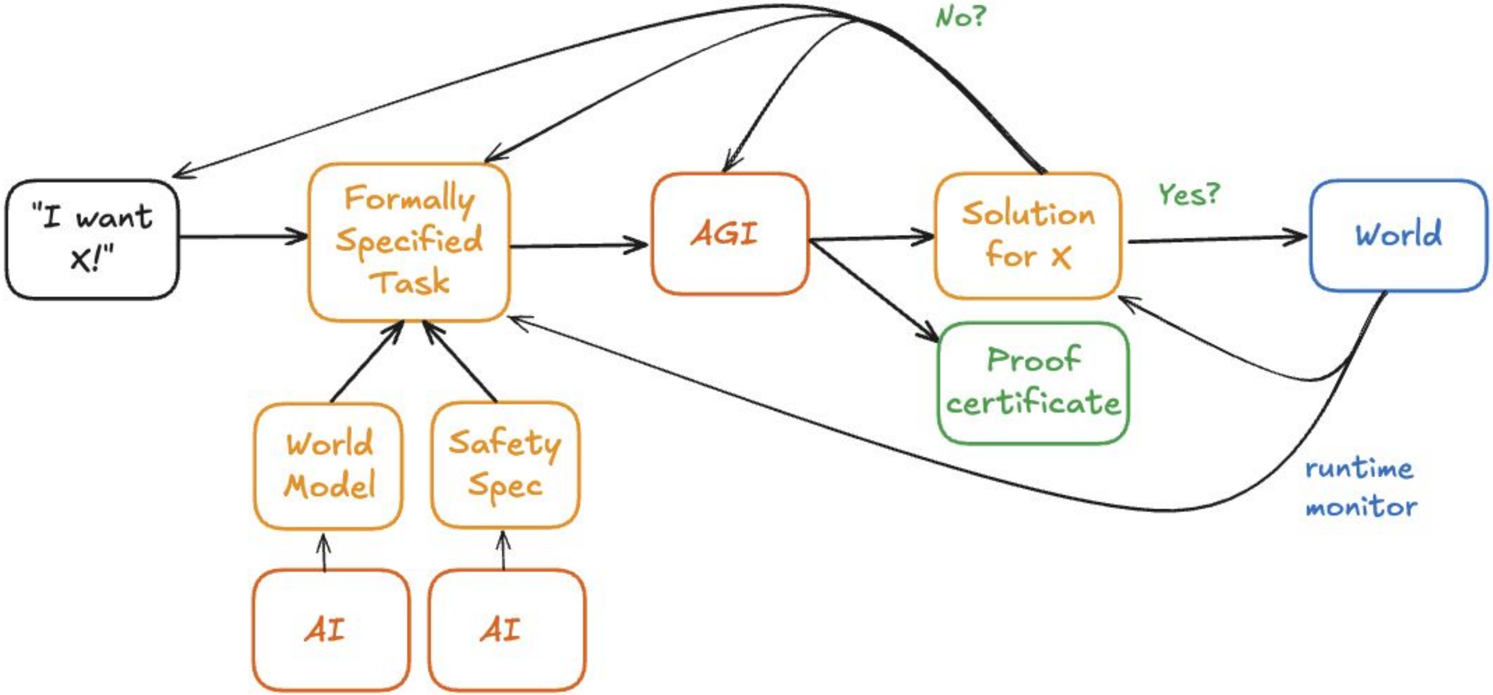
Safe AI?


"I want X!"
Formally Specified Task
AGI
Solution for X
Yes?
World
Proof certificate
World Model
Safety Spec
AI
AI

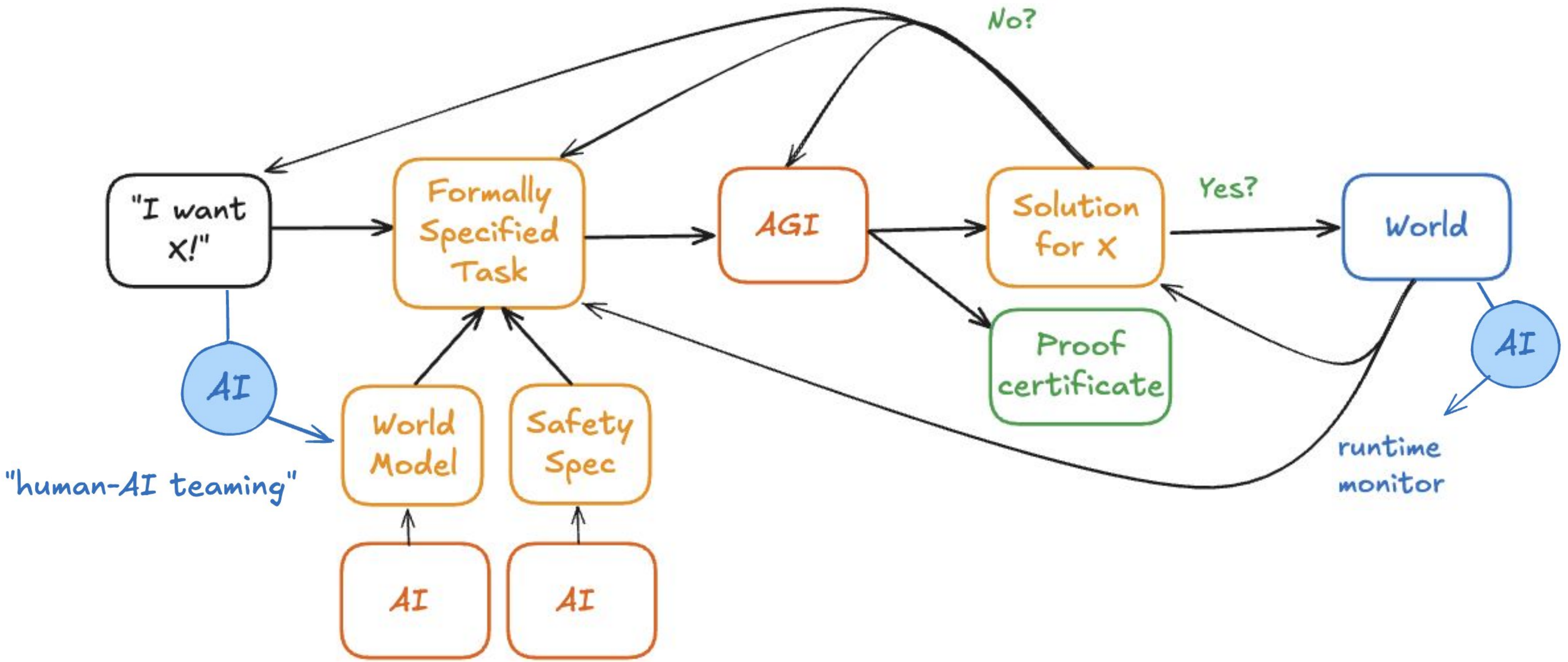
Safe AI?


No?
"I want X!"
Formally Specified Task
AGI
Solution for X
Yes?
World
Proof certificate
World Model
Safety Spec
AI
AI

Safe AI?

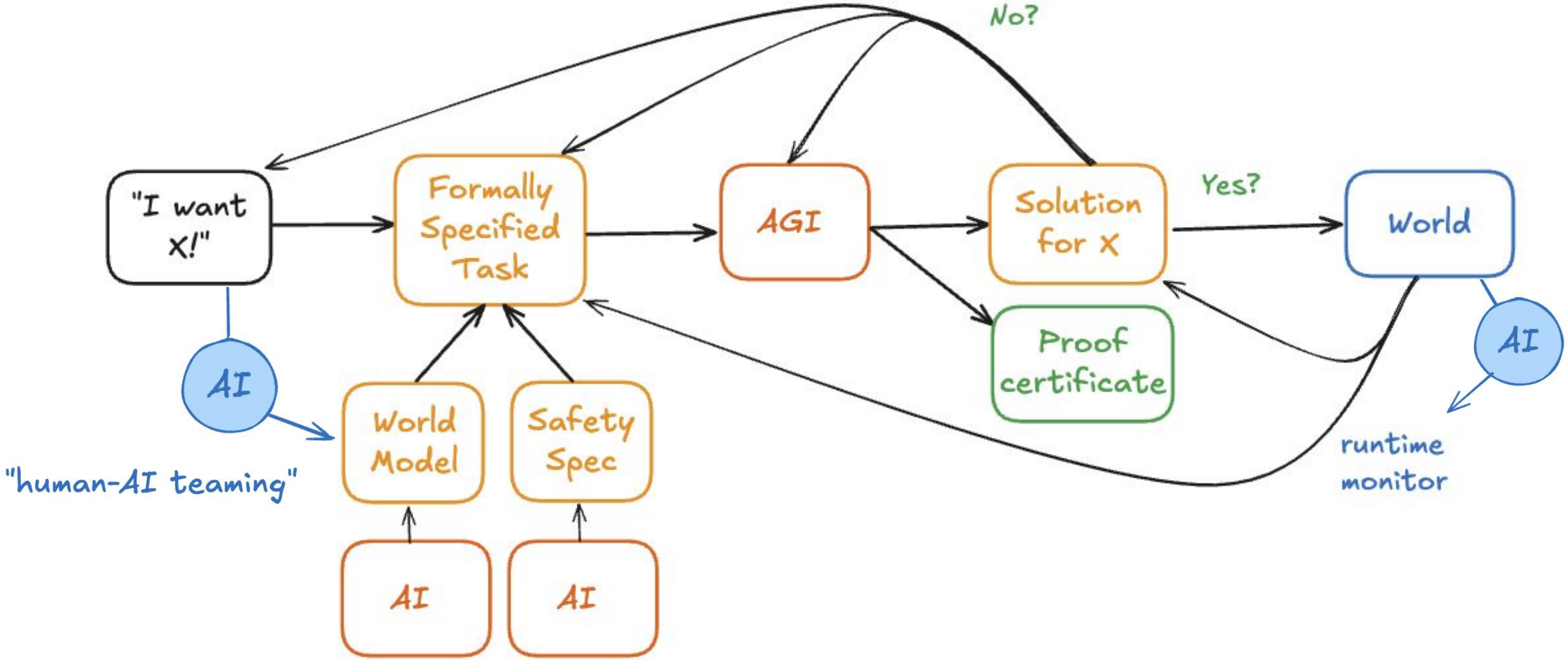

"I want X!"
Formally Specified Task
AGI
Solution for X
Yes?
World
No?
Proof certificate
World Model
Safety Spec
AI
AI
runtime monitor

Safe AI?


No?
"I want X!"
Formally Specified Task
AGI
Solution for X
Yes?
World
Proof certificate
AI
"human-AI teaming"
World Model
Safety Spec
AI
AI
AI
runtime monitor

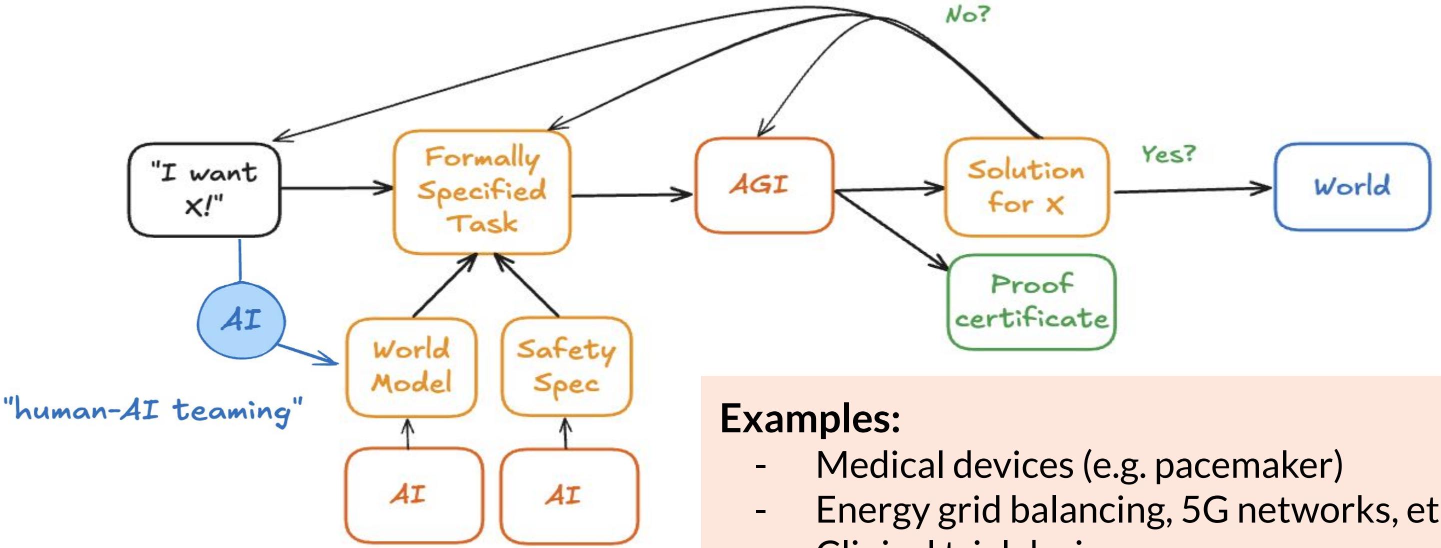
Safe AI?

See e.g. Towards Guaranteed-Safe AI (2024)



Safe AI?

See e.g. Towards Guaranteed-Safe AI (2024)



Examples:

- Medical devices (e.g. pacemaker)
- Energy grid balancing, 5G networks, etc.
- Clinical trial design
- Supply chain optimisation
- Civil engineering (predictive maintenance & planning)
- Robotic/AV control systems
- …

Call to action

There is a lot to do, and not much time!

If you are keen to build things in this space, reach out!
nora.ammann@aria.org.uk