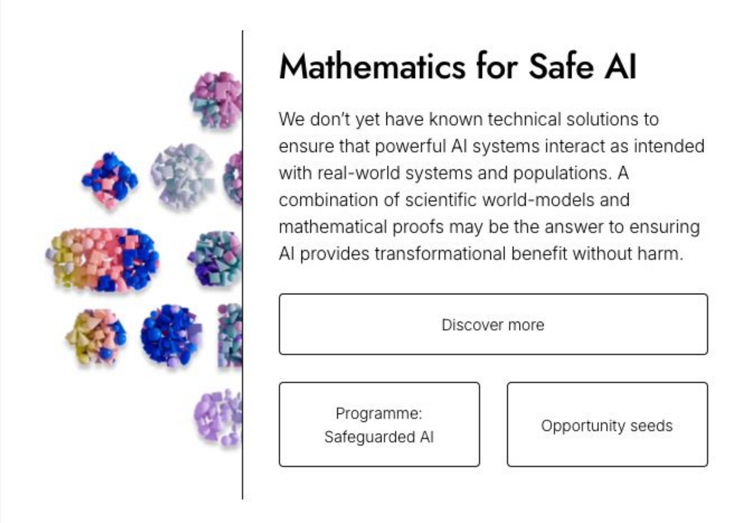
Scalable Workflows for High Assurance AI R&D

Or: How to trust the work of untrusted AIs?

Nora Ammann, 2025

ARIA



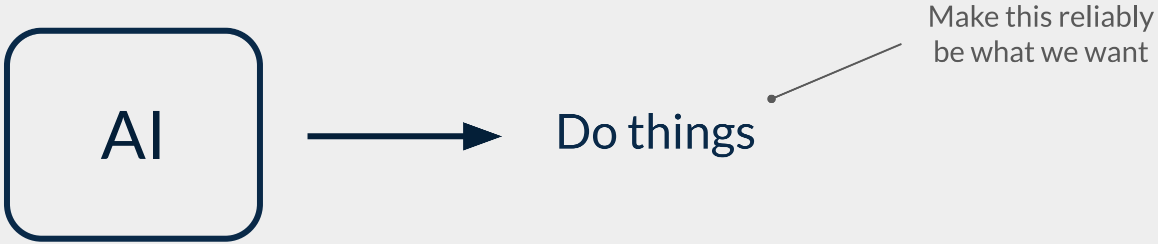
Mathematics for Safe AI

We don’t yet have known technical solutions to ensure that powerful AI systems interact as intended with real-world systems and populations. A combination of scientific world-models and mathematical proofs may be the answer to ensuring AI provides transformational benefit without harm.

My own views.

Ask questions throughout!

We haven't quite figured out the AI thing yet



Make this reliably be what we want

We haven’t quite figured out the AI thing yet

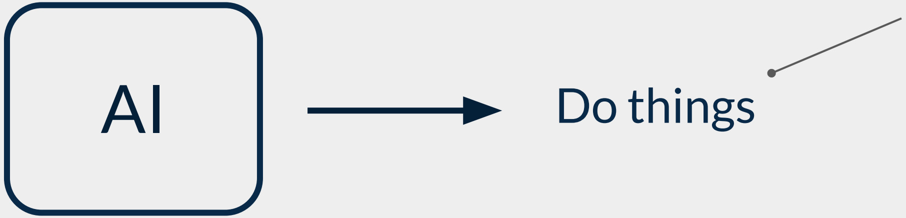
AI

Do things

Make this reliably be what we want

- They make mistakes
- It is hard to task them well
- (Sometimes we don’t exactly know what we want)
- OOD misgeneralisation

We haven’t quite figured out the AI thing yet



AI

Do things

Make this reliably be what we want

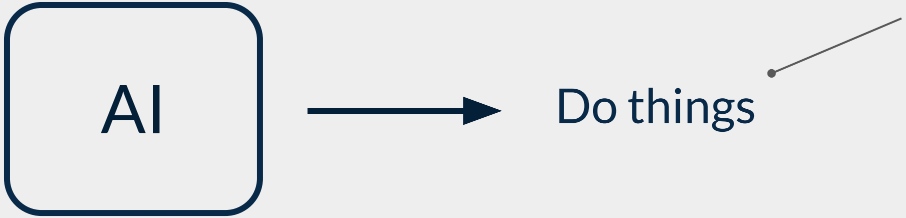
- They make mistakes
- It is hard to task them well
- (Sometimes we don’t exactly know what we want)
- OOD misgeneralisation

NB this is the case whether or not the AI is ‘friendly’

We haven’t quite figured out the AI thing yet

1. Change the internals such that you can trust the AI

2. Scaffold the (untrusted) AI such that you can trust its outputs



AI

Do things

Make this reliably be what we want

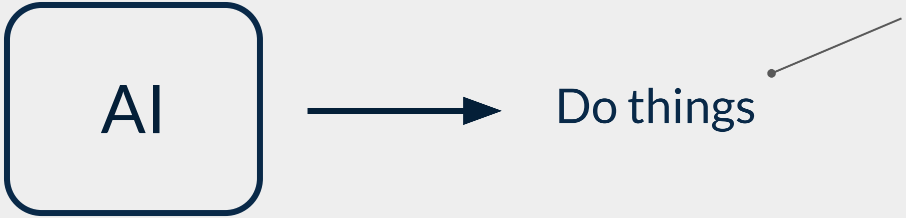
- They make mistakes
- It is hard to task them well
- (Sometimes we don’t exactly know what we want)
- OOD misgeneralisation

NB this is the case whether or not the AI is ‘friendly’

We haven’t quite figured out the AI thing yet

1 ~~Change the internals such that you can trust the AI~~
AI Alignment

2 ~~Scaffold the (untrusted) AI such that you can trust its outputs~~
AI Control / Scalable Oversight



AI

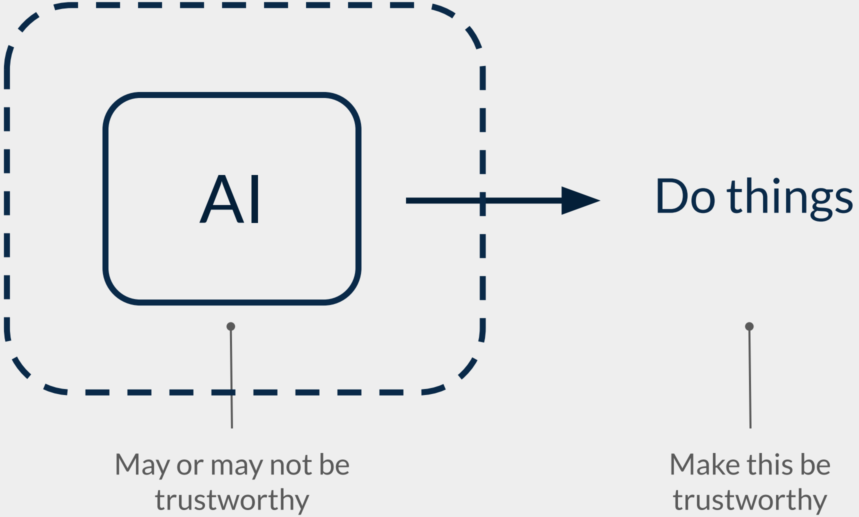
Do things

Make this reliably be what we want

- They make mistakes
- It is hard to task them well
- (Sometimes we don’t exactly know what we want)
- OOD misgeneralisation

NB this is the case whether or not the AI is ‘friendly’

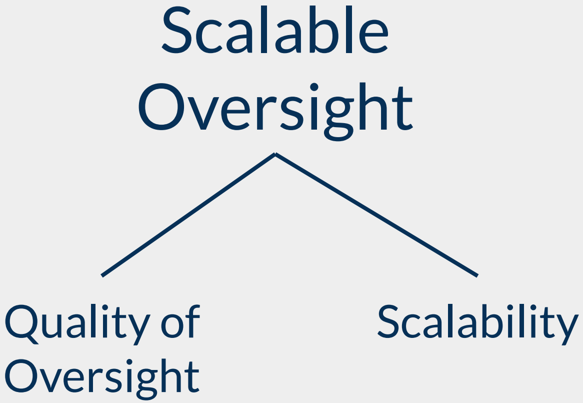
We haven’t quite figured out the AI thing yet



Scalable oversight =

Make it possible for humans to arrive at justified trust in the outputs of the AI, in a way that scales with AI capabilities.

What does ‘good’ look like?

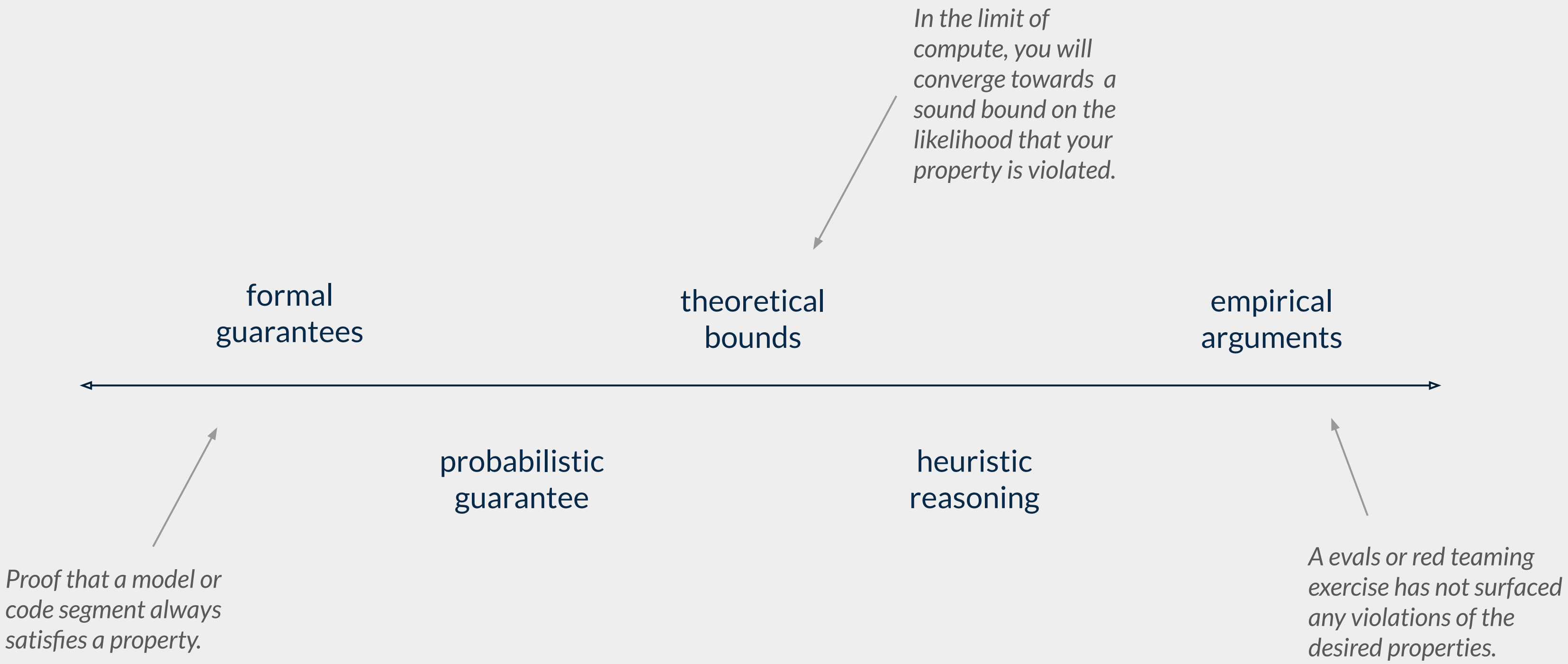


Scalable Oversight

Quality of Oversight

Scalability

A spectrum of assurance



formal guarantees

probabilistic guarantee

theoretical bounds

heuristic reasoning

empirical arguments

Proof that a model or code segment always satisfies a property.

In the limit of compute, you will converge towards a sound bound on the likelihood that your property is violated.

A evals or red teaming exercise has not surfaced any violations of the desired properties.

Scaling high-quality oversight

A general trend in R&D — away from implementation, towards oversight

E.g. “vibe coding”

- **Humans** scope a task
- **Humans** develop a solution
- **Humans** review the solution

→

- **Humans** scope a task
- **AI** develop a solution
- **Humans** review the solution

Scaling high-quality oversight

A general trend in R&D — away from implementation, towards oversight

E.g. “vibe coding”

- **Humans** scope a task
- **Humans** develop a solution
- **Humans** review the solution

→

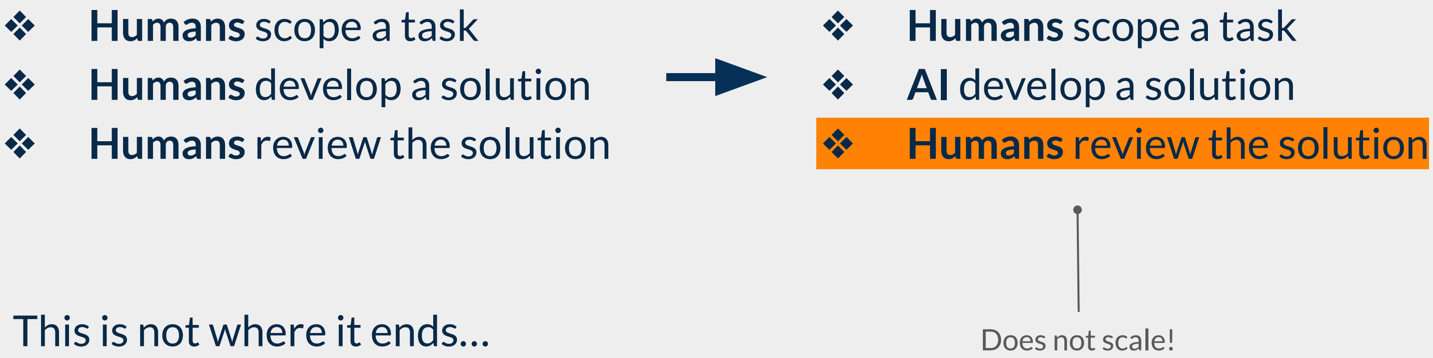
- **Humans** scope a task
- **AI** develop a solution
- **Humans** review the solution

This is not where it ends…

Scaling high-quality oversight

A general trend in R&D — away from implementation, towards oversight

E.g. “vibe coding”



- **Humans** scope a task
- **Humans** develop a solution
- **Humans** review the solution

- **Humans** scope a task
- **AI** develop a solution
- **Humans** review the solution

Does not scale!

This is not where it ends…

Scaling high-quality oversight

If we formally specify the task, we can get the AI to give us a machine-checkable proof that the solution meets those specs

A general trend in R&D — away from implementation, towards oversight

E.g. “vibe coding”

- ❖ **Humans** scope a task
- ❖ **Humans** develop a solution
- ❖ **Humans** review the solution

→

- ❖ **Humans** scope a task
- ❖ **AI** develop a solution
- ❖ **Humans** review the solution

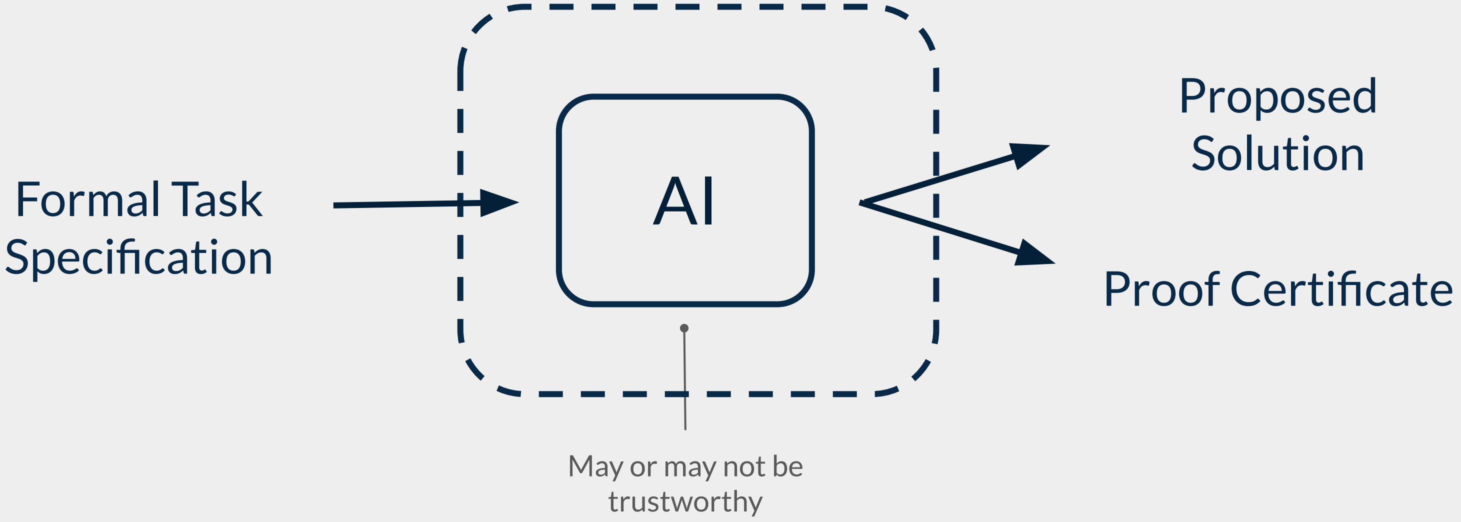
Gist: make the AI do (extra(work to make effective oversight easier for us.

This is not where it ends…

→

- ❖ **Humans** scope a task
- ❖ **AI** develop a solution
- ❖ **AI** develop a certificate of correctness

Scalable oversight



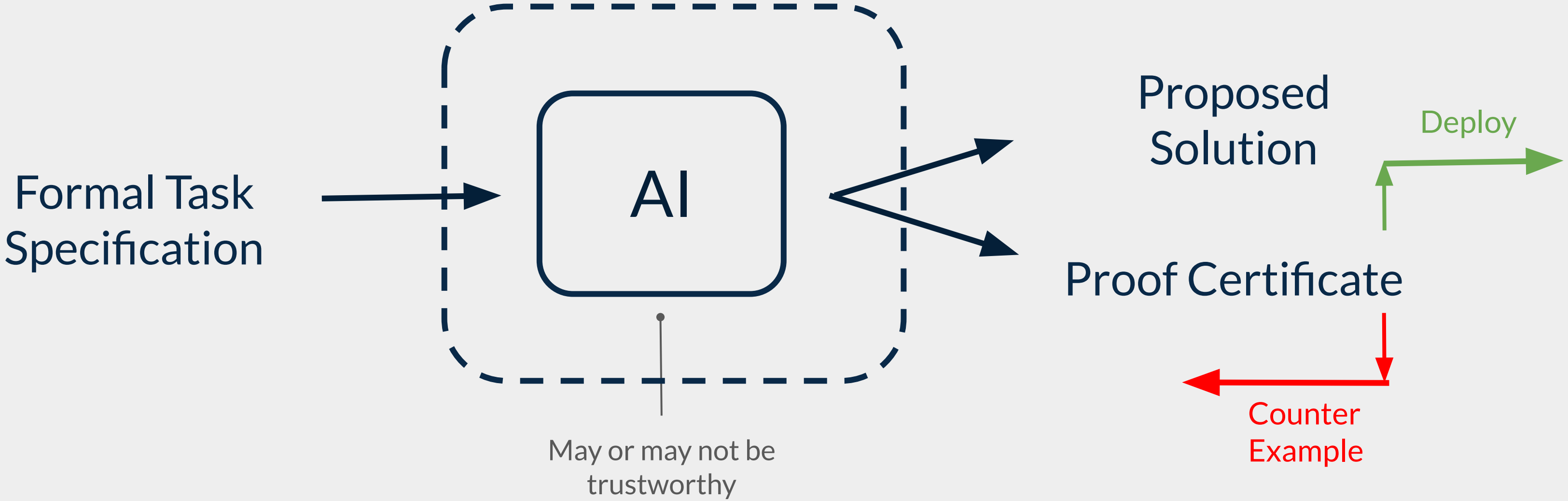
Formal Task Specification

Proposed Solution

Proof Certificate

May or may not be trustworthy

Scalable oversight



Example: Provably Secure Code

We know it’s possible.

Example: Provably Secure Code

We know it’s possible.

seL4

- Mathematically verified microkernel
- Residual defect rate <10^-9 per LOC.
- Took ~20 person years

Example: Provably Secure Code

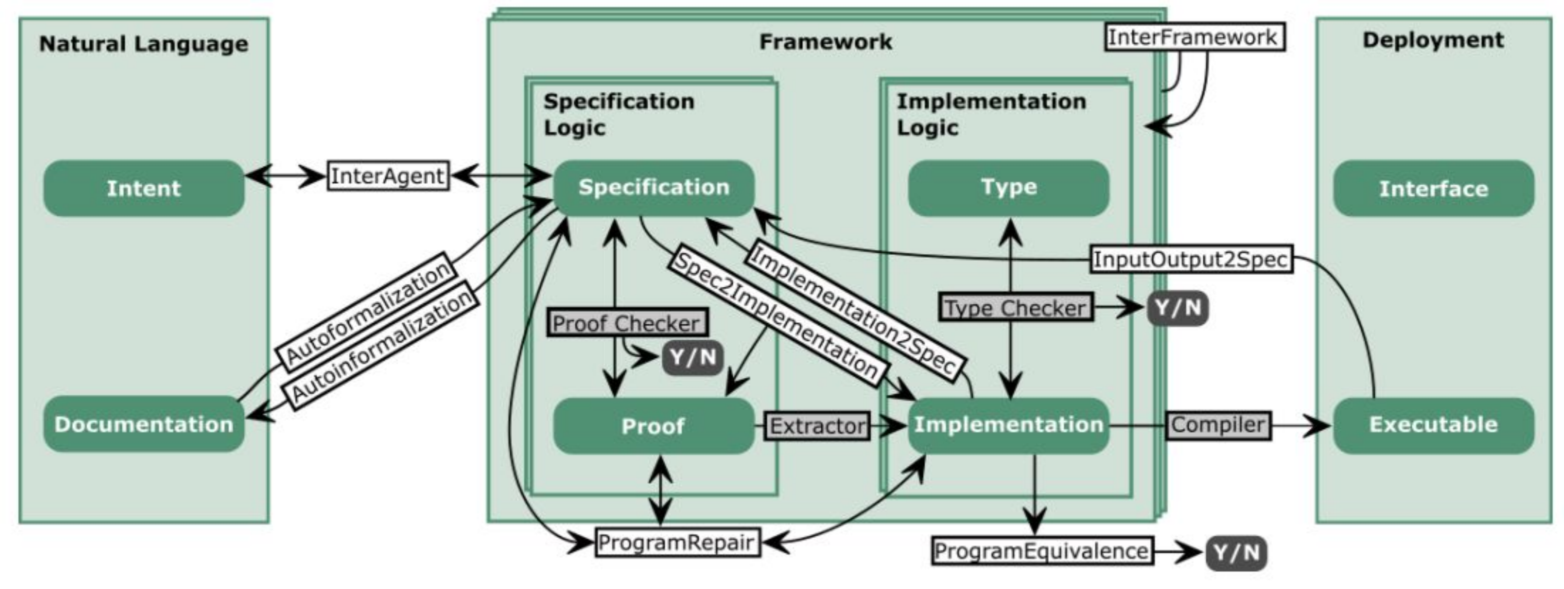
We know it’s possible.

seL4

- Mathematically verified microkernel
- Residual defect rate <10^-9 per LOC.
- Took ~20 person years

AI can reduce this work load!

Example: Provably Secure Code



Atlas Computing, 2024

Beyond code…

Energy grid balancing, supply chain management, telecom networks, pharmaceutical manufacturing, R&D planning, …

A similar workflow can be extended beyond the domain of pure software

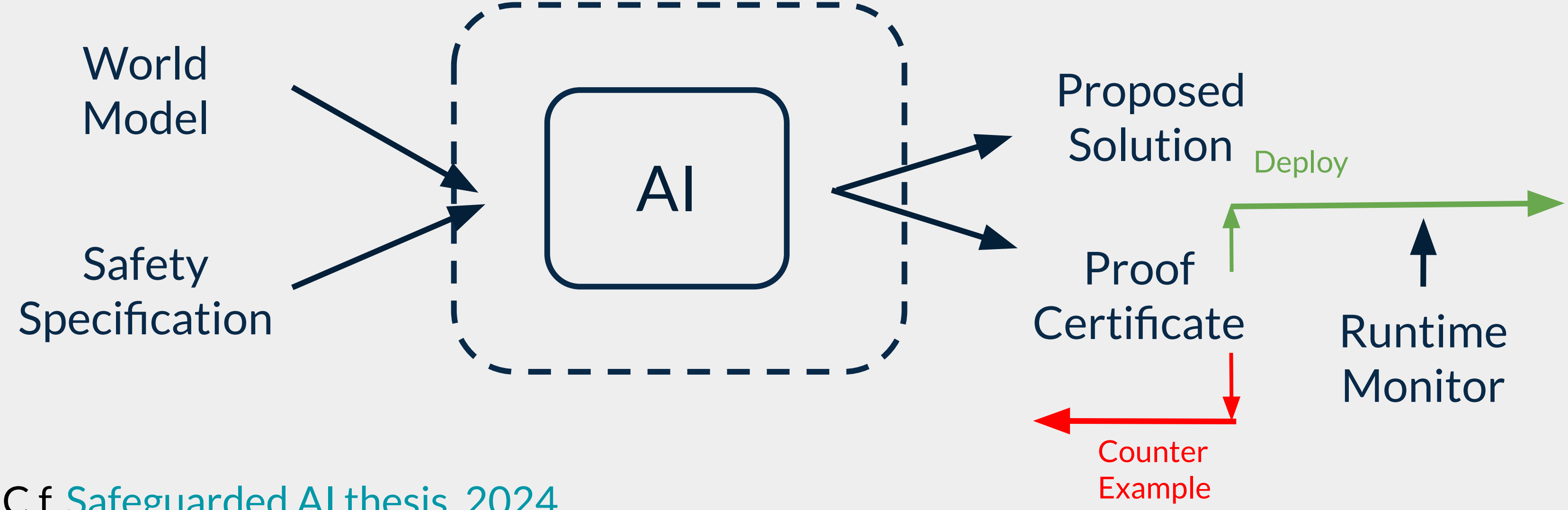
E.g. Cyber-Physical Control Systems

Beyond code...

Energy grid balancing, supply chain management, telecom networks, pharmaceutical manufacturing, R&D planning, ...

A similar workflow can be extended beyond the domain of pure software

E.g. Cyber-Physical Control Systems



C.f. Safeguarded AI thesis, 2024

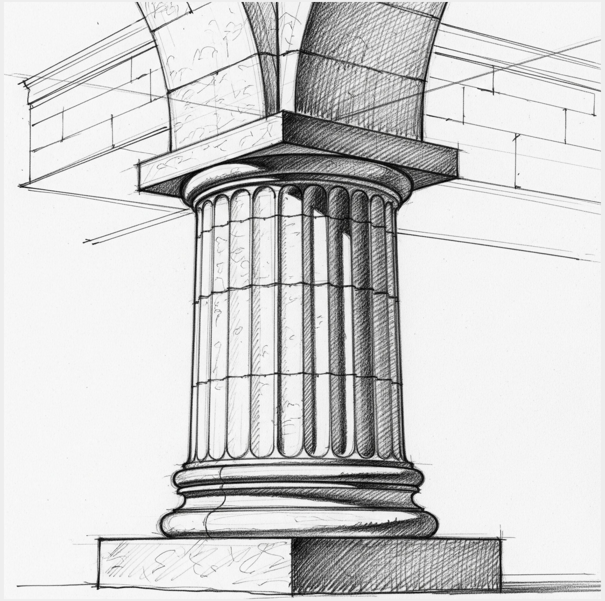
A new load bearing pillar — Spec validation

Your assurance is only as good as your specs.

Rich opportunities for novel HCI design paradigms.

Some places we can look:

- Verification & Validation (V&V) workflows
- Emerging literature on LLM coding assistance
- A new area for HCI / H-AI teaming workflows?



Ok so…

Ok so…

Getting trusted work out of untrusted AIs

Ok so…

Getting trusted work out of untrusted AIs

What can we do with that?

What is this all good for?

1. Reduces the pressure to deploy AI unsafely

What is this all good for?

1. Reduces the pressure to deploy AI unsafely
2. Helps secure critical civilisational infrastructure

What is this all good for?

1. Reduces the pressure to deploy AI unsafely
2. Helps secure critical civilisational infrastructure
3. Enables robust cooperation among agents

NB verifiable assurance is not only good for safety/security, but also for coordination.

Closing thought

- There is a rich design space for better scalable oversight schemes.
- I think we need to high for very high degrees of mathematical rigour to be able to scale high quality oversight.
- Key question: What tooling and infra to build now, assuming we will have highly competent AI agents within a year?