Mapping Architectural Solutions to AI Risk

The Atlas Computing Institute ("Atlas") is a nonprofit mapping paths to scale human review of risky AI capabilities. We're operationalizing Safeguarded AI (formerly the Open Agency Architecture) to drive creation of AI-accelerated tools that can protect against AI-accelerated risks.

Why it matters: Today's AI architectures are automating our ability to change systems without improving our review systems (e.g. code review, chemical safety review, legal review). New tools for identifying and closing loopholes is critical for a healthy civilization, which can only be scaled with a new AI architecture.

Today's AI have no guarantees, and other safety research agendas aren't even pursuing verifiability.

Society functions because we have systems to provide safety guarantees that airplanes won't crash, mobile phones won't catch fire, and medicine won't make you more sick. These guarantees are verifiable: we've stated properties objects should have and test to see if they have these properties.

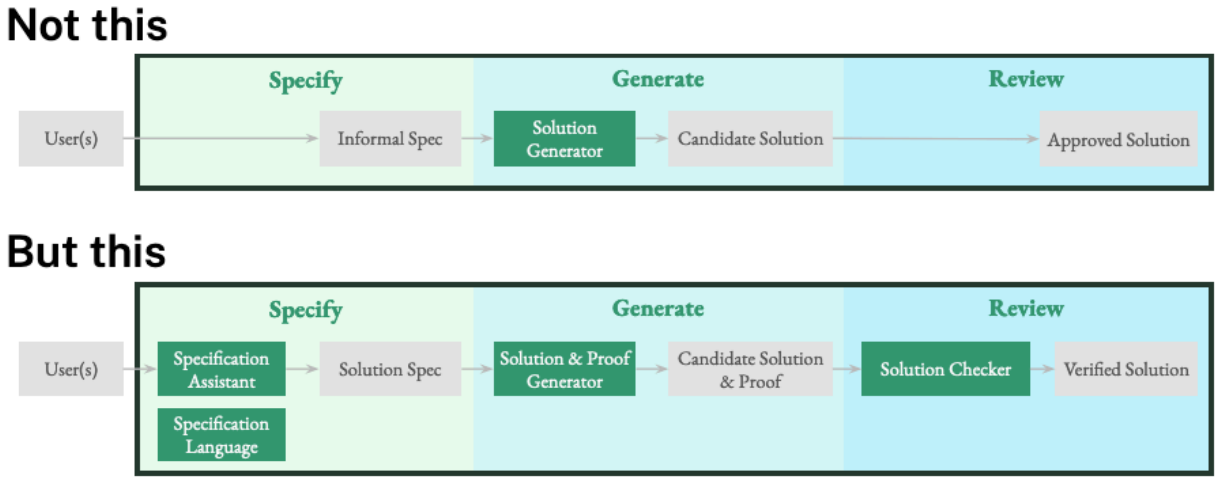
Machine learning (ML) systems probabilistically generate new output to maximize a utility function that's an external metric set in training or an internal metric during runtime. This utility function can be nudged (analogous to clicker training of zoo animals), but it's currently impossible to ensure that your system has not been compromised by biased input data or a malicious actor because there is no known way to set or understand the internal tendencies or goals of a deployed system. But we can still have verifiable ML systems.

Specification-based AI could provide objective evidence that outputs satisfy explicit constraints.

The Safeguarded AI Architecture is designed to keep humans in control by advancing capabilities by first scaling review systems. We describe this architecture as "specification-based AI" because

1. Users use AI-accelertated tools **specify** what properties they desires of outputs with the help of AI tools
2. AI-based tools **generate** candidate solutions as well as evidence that the solution satisfies the spec.
3. Users **review** the evidence and approve or reject the solution based on the evidence.

With objective evidence that outputs matche your spec (similar to how formal verification guarantees a program won't crash), you don't need to have encoded your values into the AI. Also, specifications can be a target of governance, allowing institutions to scale review with AI, rather than needing to trust the AI systems.



Operationalizing this architecture can be achieved one domain at a time.

Societies have review systems for various domains; each could be empowered or overwhelmed by generative AI.

Potential risks and tools in 5 domains (illustrative, not exhaustive)

Domain	Potential Risks	Potential Benefits
Software	Identify and exploit software vulnerabilities faster than patches are deployed	Generate provably secure software to run our computers and computer networks
Biochemistry	Generate novel bioweapons and chemical weapons from unregulated precursors.	Generate novel therapeutics and materials to solve medical, environmental, and engineering problems
Economy	Undetectably manipulate markets	Create new markets, help discover pareto-optimal market dynamics
Society	Target each citizen with personalized disinformation and cons	A world-class tutor, advisor, and therapist helping each individual to achieve their long-term goals
Regulatory	Automated detection and exploitation of legal loopholes	Tooling that improve regulatory transparency and scale adjudication systems

The primary challenges are: generating formal specifications of stakeholder objectives and solutions that satisfy these specs, which will require research and technology transfer in each domain.

The strongest reason to pursue this plan is that it is a safer approach to AI safety. Upon success, it produces safe AI capabilities. If it fails, it produces AI systems that do nothing. This is in contrast to common safety approaches in the wider AI community that risk creating opaque, malfunctioning safety systems whose dangers are hard to find until it is too late to avoid harm.

Our first milestone is demonstrating specification-based tools for cybersecurity: Atlas will create tools based on the OAA to facilitate the generation of formally specified software. Success would see advanced market commitments from governments to use tools to secure critical cyber-physical infrastructure. Read more about the opportunity for AI and formal verification here.

We will create plans and proofs-of-concept, coordinating with research groups & companies to build these products in an open ecosystem, rather than attempt to corner the market on OAA-based systems.

Go deeper on Atlas:

- A pitch-deck-style slide deck summary of Atlas Computing
- Our proposal to organize a toxicity forecasting competition to create a spec language for biochemistry

More on AI risk and the Safeguarded AI architecture

- The UK's Advanced Research and Invention Agency Programme Thesis on Safeguarded AI by Atlas advisor davidad, our executive summary of the program
- Advocacy for this category of approach from leaders in AI safety.
  - Yoshua Bengio, 2024 - Towards a Cautious Scientist AI with Convergent Safety Bounds
  - Max Tegmark, Steve Omohundro, 2023 - Provably safe systems: the only path to controllable AGI
  - Stuart Russell, 2022 - Provably Beneficial AI
- Gladstone AI's Action Plan; A plausible case for AI risks: Andrew Critch's description of risk scenarios