An architecture to build controllable AGI

The Atlas Computing Initiative ("Atlas") is a nonprofit mapping paths to scale human control over risky AI capabilities. We're using the Open Agency Architecture (OAA) to drive creation of AI-accelerated tools to protect against foreseeable AI risks.

Why it matters: Advances in AI are driving us to a world of automated loophole detection (both accidental and malicious) that could break all human review systems (e.g. code review, legal review, chemical safety review). Identifying and closing these loopholes first is critical for the well-being of civilization, and this can only be achieved effectively by scaling human governance systems.

Humanity should only allow mass-scale automation to be deployed if it's governed well.
Any safety proposal that has terrible outcomes if it misfires is a bad idea.

Potential risks and tools in 5 domains

Domain	Potential Risks	Potential Benefits
Cybersecurity	Identify and exploit software vulnerabilities faster than patches are deployed	Generate provably secure software to run our computers and computer networks
Biosecurity	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, career coach, advisor, and therapist empowering each individual to achieve their long-term goals
Regulatory	Automated detection and exploitation of legal loopholes	Tooling to scale existing adjudication systems and inform citizens

The set of domains and risks within each domain are illustrative, not exhaustive.

Other approaches to align superintelligent AI may be fundamentally flawed

Machine learning systems are trained statistically to probabilistically generate new output that maximizes some utility function, be it an external metric in training or perhaps an internal metric during runtime. Large language models are starting to exhibit some of the above risks, which arise because **there is no known way to set or understand the deployed system's internal goals or tendencies** to check the system's alignment with societal goals.

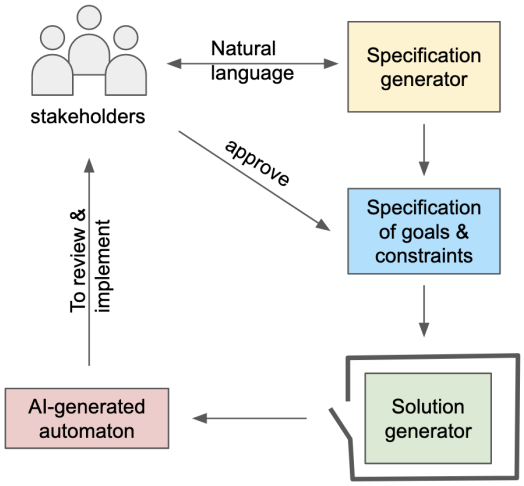
Other AI Safety research directions that expect to surpass human-level *capabilities* don't guarantee alignment with human *values*. OAA-based systems keep humans in control, scaling governance at least as quickly as capabilities.

How it works: Controllable AI makes goals & constraints explicit and objective

If you have guarantees that your AI system's behavior matches your goals and constraints, you don't need to encode your values into the AI. The OAA leverages objective and verifiable descriptions of the AI's goals (similar to how formal verification guarantees that a program won't reach undesired states). Also, controllable AI can identify trade-offs between competing requests of multiple stakeholders and let them decide, rather than abdicating the final decisions to an aligned AI.

The main components of an OAA include:

A. *Stakeholders* describing (in natural language) a solution they seek and properties of the solution;
B. *A language model* generating specifications from interaction with the stakeholders;
C. *A specification language* that objectively & verifiably describes goals & constraints;
D. *A solution generator* using machine learning to propose solutions (e.g. software patches or AI monitoring programs) and a world-simulator to check solutions against the specification.



This diagram simplifies the detailed diagram here.

The primary challenges are: generating formal specifications of stakeholder objectives and solutions that satisfy these specs, which will require significant research to solve for 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 an OAA-based system 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 (licenced by Atlas) to secure critical cyber-physical infrastructure, while we begin work on domains beyond cybersecurity.

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

Who we are: As a new nonprofit organization, Atlas Computing consists of:

Evan Miyazono, executive director, completed a PhD in Applied Physics at Caltech, going on to lead research at Protocol Labs, creating their research grants program, and leading the special projects team that created Hypercerts, Funding the Commons, gov4git, and key parts of Discourse Graphs and the initial Open Agency Architecture proposal.

Daniel Windham, software lead, previously co-led software development and usability of the STITCHES systems integration toolchain, one of the most successful DARPA programs of the decade. Previously, he worked on programming environments and HCI at Coda, Y Combinator Research, and MIT Media Lab.

Go deeper:

- Our first step, securing cyber physical infrastructure: Making software provably safe at scale
- A better understanding of the risk AI poses: Andrew Critch's description of risk scenarios
- The initial proposal of the OAA: davidad's Dec 2022 post
- The latest on the OAA: Raphaël S and Gabin's 4/23 summary; davidad's 8/23 open problems
- More specifics on what the Atlas Computing Initiative is planning: Budget and roadmap docs