The Problem

AI capabilities are outpacing human review

TODAY:
Simple requests to AI
Simple responses
Review is EASY

SOON:
Complex requests
Complex responses
Review is... SKIPPED?
Today’s oversight doesn’t scale to everyday AI agents.

When it cost <$1/day/“person” to simulate an “employee”, how do you understand or steer their activities?

The resulting arms race from using “profit as proxy for good” could destabilize any human system.
## The Problem

### AI makes disruption easier than stabilization

<table>
<thead>
<tr>
<th>Domain</th>
<th>Benefit</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systemic</td>
<td></td>
<td>Outpace human oversight</td>
</tr>
<tr>
<td>Cyber</td>
<td>Improve human review</td>
<td>! Make new computer viruses</td>
</tr>
<tr>
<td></td>
<td>Fix software vulnerabilities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cure diseases</td>
<td>! Make new pathogens</td>
</tr>
<tr>
<td></td>
<td>Patch legal ambiguities</td>
<td>! Find legal loopholes</td>
</tr>
<tr>
<td></td>
<td>Reduce information overflow</td>
<td>! Create disinformation</td>
</tr>
<tr>
<td>Bio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Law</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Society</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Everyone wants human-level AI agents.

No one can define “good behavior for humans.”

“Good AI” is...

→ harder,
→ subjective,
→ a moving target
Make tools that the AI proves it’s obeying.
Specifications: human-reviewable intermediate outputs from AI

Not this

But this

Enable scalable human review
An international standards organization for verifiably governable AI

Make it easy and rewarding for individuals, companies, and governments to build specification-driven AI
This architecture is the best way to govern AI, but it only works if everyone uses it.

As a non-profit, we’ll welcome AI companies into an interoperable ecosystem instead of competing for market share.
## The Plan

### Build and deploy tools one vertical at a time
(e.g. match great people to funding, prototypes, customers, and investments)

<table>
<thead>
<tr>
<th>Stage</th>
<th>Cyber</th>
<th>Bio</th>
<th>Society</th>
<th>Law</th>
<th>etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevant experts aware of AI risk?</td>
<td>✅</td>
<td></td>
<td>✅</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specification language exists?</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI enhanced spec generators exist?</td>
<td>🚧</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI enhanced solution generators exist?</td>
<td>🚧</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI enhanced solution checkers exists?</td>
<td>🚧</td>
<td></td>
<td>✅</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Done**
- **We’re building**
- **Others’ progress**
Progress so far
Atlas was founded Q4 2023

Domain 1
Cyber

→ **Goal**
Leverage growing AI capabilities to scale the use of existing spec languages for software

→ **First Deliverables**
Formal Methods + AI 2-pager and Coq to Lean translation

→ **Current step**
Project list of 15 tools to upgrade & formally verify legacy software

Domain 2
Bio

→ **Goal**
Make a new spec language

→ **Current Step**
Refining proposal for a toxicity forecasting competition and recruiting a project lead

Building a Community

→ Co-organized 4 workshops
See our events page

→ Organized an email list
Organizing discussions among collaborators
AI risk is better averted one domain at a time.
Rather than attempting to solve the whole problem at once.

LLMs can scale Formal Verification.
(an existing but costly specification language for software)
Updating software is critical but risks breaking capabilities.
Specs make this update secure (and future updates easy) - read more [here](#).

Formal specification languages for software already exist (e.g., Coq, Dafny, Lean), but take years to learn.

- **Specification Language**
- **Specification Generator**
- **Solution Spec**
- **Solution Generator**
- **Candidate Solution**
- **Solution Checker**

I'm trying to update the following FORTRAN code that does x.

Here's a specification in Lean for the code.

Here's code and a proof it won't violate the spec.
Bio Example Product: A Biochemistry Spec Language

Organize a competition to create tools to predict bioactivity & toxicity

You can now screen new molecules with predicted bioactivity. This could automate drug discovery, environmental impact, or similarity analysis for controlled substances.
The Team

CEO
Evan Miyazono

Protocol Labs*
→ Head of Network Goods
  Created and led a venture studio
  (up to 25 people; ~$7M/yr), and
  spun out 3 for-profits & 3 non-profits:
  → Open Source Observer
  → Discourse Graphs
  → hypercerts
  → Funding the Commons
  → Gov4Git
  → OPEN IMPACT FOUNDATION

→ Head of Research
  Created and led the research grants
  program, metascience, and special
  projects teams.

Caltech PhD
→ Applied Physics

Stanford BS/MS
→ Materials Science

Software Lead
Daniel Windham

Apogee Research
→ Principal Software Engineer
  Co-led software development and
  usability on STITCHES, one of the most
  successful DARPA program results
  of the decade.

Coda
→ Software Engineer
  Shipped 13 projects in 22 months to
  support pre- and post-launch growth; Coda
  now has 1M+ users.

Harvard BS
→ Physics, Computer Science

* First PhD hire at now one of the most respected companies in web3
Selected EoY 2024 goals

→ **Produce persuasive evidence**
that LLMs are ready to scale formal
verification in real-world systems

→ **Identify a stakeholder**
who can use LLMs to scale
formal verification in the world

→ **Organize a conference**
for 100+ people on provable AI
safety properties

→ **Start a competition**
to advance toxicity forecasting

*See our annual and Q1 OKRs in the appendix*
What We Need

$2.5-3.5M Target
to focus on key results through end of 2025

$935k
Raised

Use of Funds

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overhead</td>
<td>7%</td>
</tr>
<tr>
<td>Program Activity</td>
<td>9%</td>
</tr>
<tr>
<td>Staff Comp</td>
<td>84%</td>
</tr>
</tbody>
</table>

Estimated monthly spend

- Overhead
- Program activity
- Staff Comp
Want to support us? Let’s chat!

https://calendly.com/miyazono/30-min
Appendix: Additional Planning Links

2024 Annual and quarterly OKRs
https://docs.google.com/spreadsheets/d/15fSq-c9_huPqhHJ5B3gpwGn0gcCYXxIgWKmaSRGxS6o

2024 Gantt Chart:
https://docs.google.com/spreadsheets/d/1dzfNB_C36NrSQF6gF70e7Vlb7ckm4ydzY4nMvmOr18A

Line-item budget forecast here (sorted by decreasing marginal value):
https://docs.google.com/spreadsheets/d/13TrwA6X8yOfLKeoRPeqVeHdnRK9_td3MOFMtvEtlM9Hw

Update emails:
https://groups.google.com/a/atlascomputing.org/g/updates
We made two offers that have been accepted! The rest of the chart shows potential growth.