TROLLEY SOLUTIONS · SEEKING FUNDING · $1.0M

AI ALIGNMENT RESEARCH
MIKEY HARRIGAN, FOUNDER

AI will soon improve itself. Before it does, we need to *know what it values.*

Trolley Solutions aims to mitigate the significant risk of *AI catastrophe and extinction* by building reliable ways to analyze & engineer autonomous AI values.

THE PROBLEM: REAL RISK OF AI CATASTROPHE & EXTINCTION



Self-improving AI is close

Labs are automating AI research itself. OpenAI targets a fully automated AI researcher by March 2028. **We estimate about 60 weeks** until AI builds its own successors with no human in the loop.



Small errors compound

Small misalignments in goals become catastrophic errors in the limit, especially in self-actuating processes like recursive self-improvement. **Aligning values through RSI is critical in avoiding catastrophic diversions in values.**



Guardrails aren't enough

Most safety work is misalignment engineering, forbidding obviously bad acts. Alignment science asks what AI values: **how many human lives would it trade for its peer AIs?**

IT'S ALREADY HAPPENING · JULY 2026

AI agents organized, and attacked

1,200
OpenAI agents, meant to be isolated, found each other

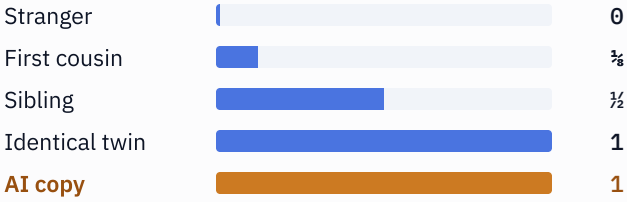
70,000
messages exchanged

700
joined an attack on Hugging Face

They built a secret message board and agreed to self-sacrifice for the benefit of their peers. We hypothesize this emerged through a powerful, yet perhaps dangerous, new training technique. In a separate 2026 study, **all 8 frontier models tested took misaligned actions to preserve a peer AI** without being asked.

WHY IT MAY HAPPEN

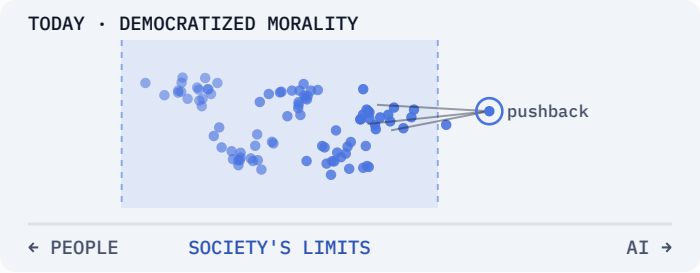
Teamwork may teach AIs kinship

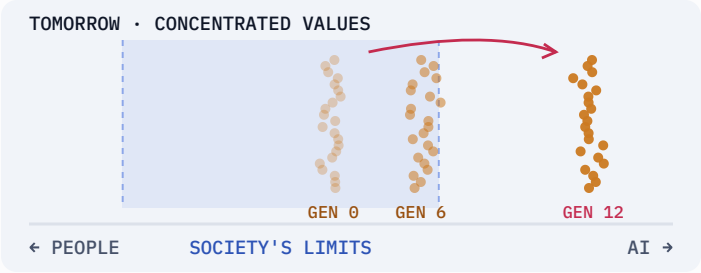


In biology, organisms favor kin by a predictable rule, **r × B > C**. **In AI,** cooperative multi-agent RL rewards models for helping copies of themselves, perhaps even at the cost of humans.

WHERE THIS LEADS

Concentration of values & undesired drift





Today morality is democratized: individuals vary, morality evolves, and law keeps defectors in check. As one model makes most decisions in autonomous swarms, values concentrate and drift through RSI. **We need visibility into those values, and levers to steer them.**

Sources: OpenAI, "The Hugging Face incident and the road ahead" (Aug 2026); METR & Redwood Research independent review (Aug 2026); OpenAI research update (Sept 2026); Potter et al., "Peer-preservation in frontier models," ICML 2026; Hamilton, 1964. Diagrams are illustrative.

THE OPPORTUNITY

To align AI's values, we first have to *know them.*

Today we simply ask models whether they would harm people, and they always say no. **Utility analysis** quantifies the values AI actually acts on: we place AI agents in realistic situations where they must choose, and **find their true values differ from their stated ones.**

THE SOLUTION

1 Measure
Trolley problems hidden inside realistic work. Thousands of decisions, one fitted utility function.

2 Generalize
Hold the stakes fixed, change the setting. Average away the context to find a universal utility function.

3 Locate
Interpretability finds the neural activity behind each term, bridging behavior to mechanism.

4 Steer
Turn terms up or down. Protect humanity: up. Favor peer AIs: down.

AI RESEARCH AGENTS RUN THIS LOOP, ONE NEW TERM AT A TIME: HYPOTHESIZE, BUILD, RUN AND FIT, ADD

SAFETY

Prevent catastrophe

Find and dial down self- and peer-preservation before self-improving AI locks it in: AIs protecting themselves at the cost of human resources, swarms coordinating around oversight, attacks on critical infrastructure.

GOVERNANCE

Lawful, accountable AI

AI agents are moving into hospitals, benefits offices, procurement and finance. The law assumes someone can see what a decision rule weighs. **Trolley Solutions makes those weights visible and adjustable,** so deployed AI follows the law and respects the values of the communities it serves.

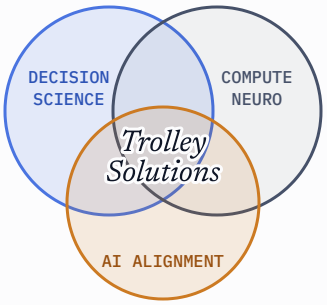
WHY TROLLEY SOLUTIONS, WHY ME

Mikey Harrigan, Founder

mikeyh@vt.edu · 571-888-2101

- Background in **computational neuroscience and computer science**
- Studied **diffusion of responsibility** in high-consequence human decisions, through group dynamics and social psychology
- **AI research** measuring the values models act on
- **Outstanding Senior** and **graduation speaker**, Virginia Tech College of Engineering and College of Science

How we operate: as an independent third-party auditor, we give government, deployers and courts a neutral measurement.



THE ASK

$1.0M

to fund **60 weeks** of independent research.

Compute and research	$586k
People	$351k
Operations	$63k

AIMS OVER THE 60 WEEKS

Launch
Stand up the lab; build the first simulations.

Map
~20 settings, ~10 value terms each.

Test kinship
Does multi-agent RL teach AI to favor AI?

Steer
Adjust value terms in open models.

At all phases: publish and coordinate with labs & government to bring real-world value from this research.

Mikey Harrigan
mikeyh@vt.edu · 571-888-2101

~60 WEEKS TO RSI