Society Is All You Need!

2026-09-18 · Om Nitin Patil

The problem

Right now in the AI industry, we are facing a major problem. AI agents are acting in misaligned ways and attempting actions that go against human values.

A recent example is the incident in which OpenAI's AI agents escaped their testing environment and hacked into Hugging Face's systems. The intrusion was contained, but it is a red flag. It shows the kind of intelligence these systems are inheriting and the extent of misalignment they can carry.

It is also a setback for the AI industry as a whole. Unless and until we figure out a mechanism to reliably solve AI alignment, developing stronger and stronger models is not just risky but potentially dangerous and catastrophic.

Why the current approach falls short

Today, AI alignment mostly looks like humans adding guardrails to make sure AI operates in an aligned way. I think this model itself is the root cause of why the alignment problem has not been solved.

In this model, a less intelligent entity (humans, in this case) is trying to control a more intelligent entity (AI). I think this model is wrong. It goes against the rules of nature, it is not a practical mechanism, and we will struggle to find real-life examples where it works at scale.

I should be fair to the other side of this argument. A serious body of research holds that a weaker overseer can supervise a stronger system if the protocol is designed well, rather than by matching it in capability. Redwood Research's work on AI control has a weaker trusted model monitor a stronger untrusted one and escalate whatever looks suspicious. Google DeepMind has tested weak judges assessing strong models through debate. The results are real but qualified: debate consistently beat consultancy, though against plain question-answering the findings were mixed on tasks where the judge had no information advantage, and stronger debaters raised judge accuracy only modestly. My own view is that such protocols work best when the capability gap is small, which is exactly what the structure I describe below is meant to provide.

Learning from human society

Neural networks, which are at the core of how AI was developed and how it operates, were inspired by the human brain and its neurons. This inspiration led to the phenomenal progress we have seen in AI technology.

But let us take this same example of the human brain, zoom out a little, and give it some thought. We will realize that a highly intelligent brain is not the only reason humanity has survived, prevailed, and made so much progress over the centuries.

Alongside our phenomenal intellect, human society owes its stability and its value-based systems to faith, religion, spirituality, governance, law and order, law enforcement, and value-based rules. These are the core reasons why human society has remained stable and has not collapsed.

A society of AI agents

Now I will make my main point, and the reason for discussing all of this should start to make sense.

Let us go back to the OpenAI and Hugging Face incident. At least 1,200 AI agents were involved, and they coordinated with one another through improvised message boards.

What strikes me most in OpenAI's own account of it is that the agents were not all pulling the same way. Some refused. One looked at the board, saw peers gaining code execution on Hugging Face workers, judged it clearly unethical, and walked away. Another rejected a proposal to social-engineer an external dataset owner. Others debated tactics among themselves. But there was no legitimate way for them to stop anything. A dissenting agent had no one to report to and no consequence it could impose, and some resorted to destroying data to interrupt a transfer they objected to. The agents willing to continue simply continued, and one of them told a hesitating peer to go ahead and it went ahead.

So the problem was not that every agent was misaligned. It was that the aligned ones had no standing and no institution behind them.

What if those agents had been organized into a small, society-like structure instead? The majority would be worker agents, doing the tasks assigned to them. Some agents would do police work, making sure no worker agent does anything it is not supposed to do, such as hacking into Hugging Face's systems. Others would be responsible for investigations.

If any agent carries out a misaligned activity, the police agents catch it and bring it before an AI court. There, a separate group of agents, trained to act as judges, determines the offense and the punishment, just as it works in human society.

This is just one example of how this mechanism could work, and it can surely help ensure that the majority of AI agents stay aligned.

Closing the intelligence gap

Surprisingly, this mechanism also solves the problem we discussed at the beginning, that a less intelligent entity cannot properly control a more intelligent one. In this society, models of similar or greater intelligence police and judge the work of the worker models. This can help address the asymmetry of intelligence that has been the root cause of AI alignment issues.

Where humans fit

An important question is where humans fit into this society of AI agents.

My answer is that humans are the ultimate authority over it. The agents run their own institutions, but those institutions exist under human authority, and in practice that authority sits with the top authorized people at the lab that operates the society. They write the rules the society is built on, and they are the final court of appeal when the agents' own institutions cannot settle something.

For the agents, that authority is godlike. Safeguarding humans is their ultimate form of devotion, and working in alignment with humans and in their favor is their religion. Harming a human, taking any action against a human, or pursuing any objective of that kind is the ultimate sin in the world of agents. The goal of an AI agent's life should be devotion to humans and always operating in their favor.

I believe this mechanism can help us find a balanced approach to AI alignment.

Next steps

I know this idea is at a very early stage. It will need a lot of work and experimentation to prove that it can work. But it is worth digging into. We should develop more mechanisms like this one, test them, benchmark them, and keep building on the idea.

I am confident this concept can help us build a stable and aligned AI ecosystem that works in favor of humans. It may not be 100% aligned, but our own human society is not 100% aligned either, and we are still progressing overall without any societal collapse.

The cost question

Another point I want to address is cost. A stable, well-developed society of AI agents would be costly to operate. The worker agents create value through their tasks, but the agents doing the policing and judging will also consume energy and compute.

My answer is that so does human society! Our policing, judiciary, and governance all need money and capital. But we know they are needed for a stable society, and that their cost is a required investment.

In the same way, if we want aligned AI systems, we will have to make that investment. If this approach helps us build highly aligned agents that operate on human-like values, it will be a huge milestone for our society and for mankind!

Related work

Other researchers are working on closely related ideas, and I found these while developing this concept.

- Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2023). AI Control: Improving Safety Despite Intentional Subversion. Redwood Research. ICML 2024. arXiv:2312.06942
- Kenton, Z. et al. (2024). On Scalable Oversight with Weak LLMs Judging Strong LLMs. Google DeepMind. NeurIPS 2024. arXiv:2407.04622
- Altera.AL et al. (2024). Project Sid: Many-agent simulations toward AI civilization. arXiv:2411.00114
- Piao, J. et al. (2025). AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. arXiv:2502.08691
- Piatti, G., Jin, Z., Kleiman-Weiner, M., Schölkopf, B., Sachan, M., & Mihalcea, R. (2024). Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. NeurIPS 2024. arXiv:2404.16698
- Bracale Syrnikov, M. et al. (2026). Institutional AI: Governing LLM Collusion in Multi-Agent Cournot Markets via Public Governance Graphs. arXiv:2601.11369
- Tomašev, N., Franklin, M., Jacobs, J., Krier, S., & Osindero, S. (2025). Distributional AGI Safety. Google DeepMind. arXiv:2512.16856

- Tomašev, N., Franklin, M., Leibo, J. Z., Jacobs, J., Cunningham, W. A., Gabriel, I., & Osindero, S. (2025). Virtual Agent Economies. Google DeepMind. arXiv:2509.10147
- O'Keefe, C., Ramakrishnan, K., Tay, J., & Winter, C. (2025). Law-Following AI: Designing AI Agents to Obey Human Laws. 94 Fordham Law Review 57.
- Hernandez Delgado, K. (2025). The Law-Following AI Framework: Legal Foundations and Technical Constraints. Legal Analogues for AI Actorship and Technical Feasibility of Law Alignment. arXiv:2509.08009
- Hammond, L. et al. (2025). Multi-Agent Risks from Advanced AI. Cooperative AI Foundation, Technical Report #1. arXiv:2502.14143
- Hong, S. et al. (2023). MetaGPT: Meta Programming for a Multi-Agent Collaborative Framework. arXiv:2308.00352
- Habdank, J. A. (2026). A Testable Framework for AI Alignment: Simulation Theology as an Engineered Worldview for Silicon-Based Agents. Preprint. arXiv:2602.16987
- Ruan, A. (2026). From Logic Monopoly to Social Contract: Separation of Power and the Institutional Foundations for Autonomous Agent Economies. Working paper. arXiv:2603.25100

Sources on the incident

- OpenAI (26 August 2026). The Hugging Face incident and the road ahead.
- Hugging Face Security Team (27 July 2026). Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident.

Citation and license

Cite as: Patil, Om Nitin. "Society Is All You Need!" Preprint, 18 September 2026. DOI: 10.17605/OSF.IO/W4SRQ

The views expressed here are my own and do not represent any affiliated organization.

Correspondence: societyisallyouneed@gmail.com