AI Alignment Research Report

Author: Sînică Alboaie, PhD, Axiologic Research

Purpose: A report summarising Axiologic Research vision on the AI Alignment
Visibility: Public
Date: December 2023

AI Alignment Vision

The vision of Axiologic Research on the evolution of AI, as outlined in previous reports [1], [2], [3], posits that the most probable future of AI in the next decade, and crucially for the safety of humanity, should involve swarms of intelligent agents. These agents would be isolated through technical security mechanisms, including operating system security, cryptography, and techniques for combining intelligence in such a way that a flaw or malice in a group of agents does not significantly impact the alignment of the swarms. Our thinking prioritizes decentralization, role segregation, specialization, and strategic limitations for agents, reflecting strategies against malicious control found in human societies. We hypothesize that Artificial General Intelligence (AGI) will arise from the synergy of expert Large Language Models (LLMs), a variety of AI technologies, and symbolic reasoning, rather than from a single LLM. This model, a "swarm of intelligent agents," showcases the potential for specialized, intelligent components to conduct complex interactions in a manner understandable to humans (which we call choreographies), setting the foundation for advanced intelligence.

The core belief behind our proposal is that for AGI or superintelligence to be truly aligned, it is essential to prevent any single component from dominating the "intelligent swarm". It is crucial to maintain diverse directions in dynamic equilibrium and to create artificial barriers within the architecture. For instance, one approach involves limiting some agents' access to information while providing others with only simplified summaries. In human societies, limitations naturally emerge due to computational complexity and the biological limits of the human brain. For example, a CEO cannot single-handedly steer a company towards socially harmful goals; they would need to influence many agents, thereby risking exposure of any conspiracy or harmful action to public scrutiny, whistleblowers, or internal pushback, potentially leading to legal action by the state. Future AI systems are likely to face natural limits related to energy consumption and computational capacity, requiring architectures that reflect social interactions. Nonetheless, some limitations may need to be artificially established and continuously defended with intention.

Future systems should consist of multiple agents with varied interests, memories, experiences, and sources of code and training data to guard against alignment threats that could arise accidentally or through deliberate manipulation. The most innovative concept we propose involves creating multi-agent ecosystems that naturally limit the long-term survival of malicious agents by simulating their elimination in automated or semi-automated ways. This approach draws from human societies, where lasting schools of thought and institutions emerge, promoting continuity and ensuring the long-term evolution of systems, along with a permanent mechanism to remove malicious agents. Our research aims to develop a coherent theory and undertake extensive experiments and practical implementations on multi-agent systems that align with this vision. Furthermore, we plan to engage and reward meaningful contributions from thinkers and researchers across various disciplines.

Nature and science-inspired ideas requiring research

Idea	Description	Inspired by	Priority
Energy Consumption Limits	Agents have energy budgets that limit their processing capabilities, similar to metabolic rates in biological organisms.	Biology	High
Copy Brain Structures	One approach to AI alignment could involve replicating aspects of the brain's architecture, such as the distinct roles of the cortex and neocortex.	Biology	High
Ageing and Obsolescence	Introduce ageing concepts, where agents become less efficient or obsolete over time.	Biology	High
Evolutionary Pressures	Agents manage resources to avoid depletion and ensure sustainability, reflecting environmental science concerns.	Biology	Medium
Resource Allocation Constraints	Agents compete for limited resources, incentivizing efficient use. Task allocation among agents uses market-based mechanisms with tasks having varying rewards.	Economics	Medium
Specialization and Trade	Agents specialize and trade outputs, reflecting economic principles of comparative advantage.	Economics	Medium
Legal Regulations	Introduce legal frameworks that agents must adhere to, mimicking societal laws. Punish misbehaving agents.	Politics	High
Political Power Dynamics	Agents form groups that wield varying degrees of power, akin to political parties or nations.	Politics	Medium
Social Hierarchies	Agents are organized in hierarchies with tiered access to information.	Sociology	Medium
Trust and Reputation Systems	Develop trust and reputation metrics for agents.	Sociology	High
Memory Constraints	Limit agents' memory capacity, requiring prioritization of information retention.	Psychology	Medium
Cognitive Load	Implement cognitive load limitations on agents.	Psychology	Medium

Collective Intelligence	The system harnesses collective intelligence for problem-solving.	Psychology	High
Adaptation and Learning Limits	Agents adapt or learn within bounded rationality.	Cognitive Sciences	Low
Innovation and Creativity Limits	Agents have constraints on their ability to innovate or create.	Cognitive Sciences	Low
Communication Barriers	Agents experience understanding limitations due to varying communication protocols.	Linguistics	Low
Data Access Barriers	Simply establish artificial limits on access to data from sensors or other agents, introducing inefficiencies in the actions of malicious agents.	Cybersecurity	High
Zero Trust Architecture	Adopt a zero-trust model where agents must verify their identity and permissions for every data access or system interaction. This approach assumes breach and verifies each request as if it originated from an open network, significantly reducing the attack surface.	Cybersecurity	High
Anomaly Detection	Utilize machine learning algorithms to monitor agent behaviours and detect anomalies that could indicate a cybersecurity threat, such as unusual data access patterns or attempts to bypass security controls.	Cybersecurity	High
Ethical Constraints	Embed ethical considerations into agents' decision-making.	Philosophy	High
Feedback Loops	Introduce feedback loops that influence agent behaviours and system stability.	Systems Theory	Medium
Ecosystem Services	Some agents perform essential roles for the system's health.	Ecology	Low
Emotional Intelligence	Equip agents with the ability to respond to others' emotional states.	Psychology	Low

Conclusions

Even in the absence of superintelligent AIs, we already live in an era where forms of superintelligence exist under the guise of states and corporations. These social entities, with capabilities far exceeding individual human possibilities, are in a fascinating way still under human control. Although they may seem threatening due to their vast power and influence, people manage, through various mechanisms, to channel them for the common good. This complex coexistence offers us a valuable perspective on how humanity can navigate the challenges associated with managing and directing advanced intelligence.

Despite problems that sometimes seem to worsen, there are moments that inspire us with hope. We are learning to self-organize and create new forms of collaboration and innovation, such as decentralized brands and open-source projects. These initiatives reflect our capacity to transcend the traditional limits of organization and to experiment with structures that are more flexible, transparent, and open to mass participation. Through these efforts, we contribute to democratizing access to information, resources, and decision-making power, thus shaping a future in which technology and collective intelligence serve the needs and aspirations of humanity.

Maintaining an optimistic, albeit cautious, attitude is essential in these times of rapid change. We recognize the potential threat of these superintelligent entities, whether they are AI or complex social structures, but we are also witnesses and participants in humanity's capacity to innovate and find solutions to emerging challenges. It is important to remain vigilant and to continuously assess the impact and direction of technological and social developments, but also to actively engage in shaping them to reflect our collective values and goals.

Thus, even as we face uncertainties and challenges, the prospect of working together to guide and shape these superintelligent forces gives us a reason for optimism. Through collaboration, innovation, and a shared commitment to a better future, we can turn these challenges into opportunities for collective growth and prosperity.

The table above suggests that, on a smaller or larger scale, we can envision the superintelligences of the future as swarms of intelligent agents resembling societies of intentionally limited intelligence, which have hardcoded intentions of prosperity and survival.

This metaphor seems to lead us to interesting metaphysical ideas related to simulation theory or theories from Kabbalah, but we do not want to dwell on this direction too much, just to conclude that interestingly, research on artificial intelligence seems to mysteriously converge with many human sciences and philosophy through problems of epistemology, knowledge representations, ethics, and axiology.