AXIOLOGIC

AI Alignment Research Report

Author: Sînică Alboaie, PhD, Axiologic Research
Purpose: A report summarising Axiologic Research vision on the AI Alignment
Visibility: Public
Date: February 2024 (Version 2 Update)

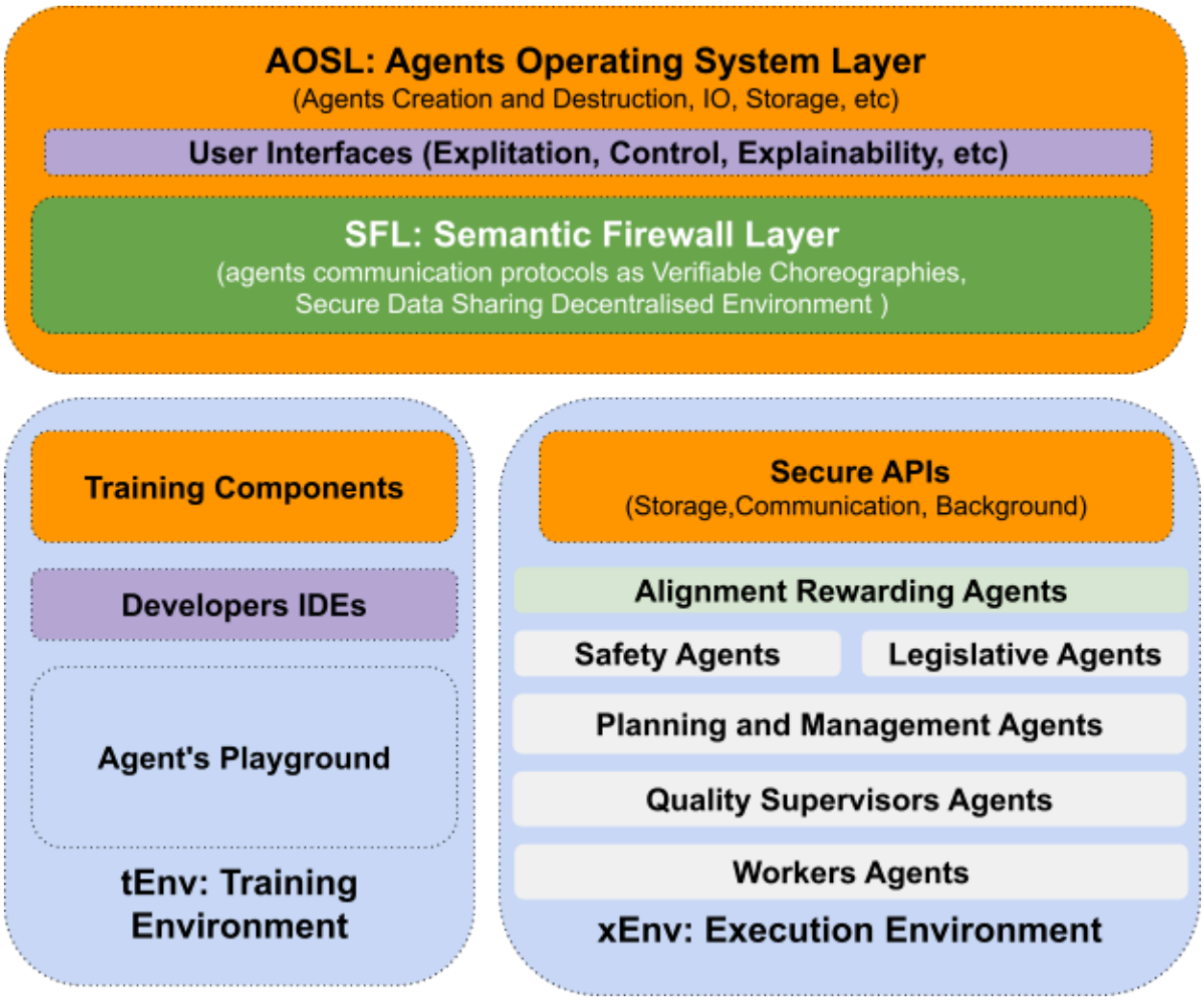
AI Alignment Vision

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.

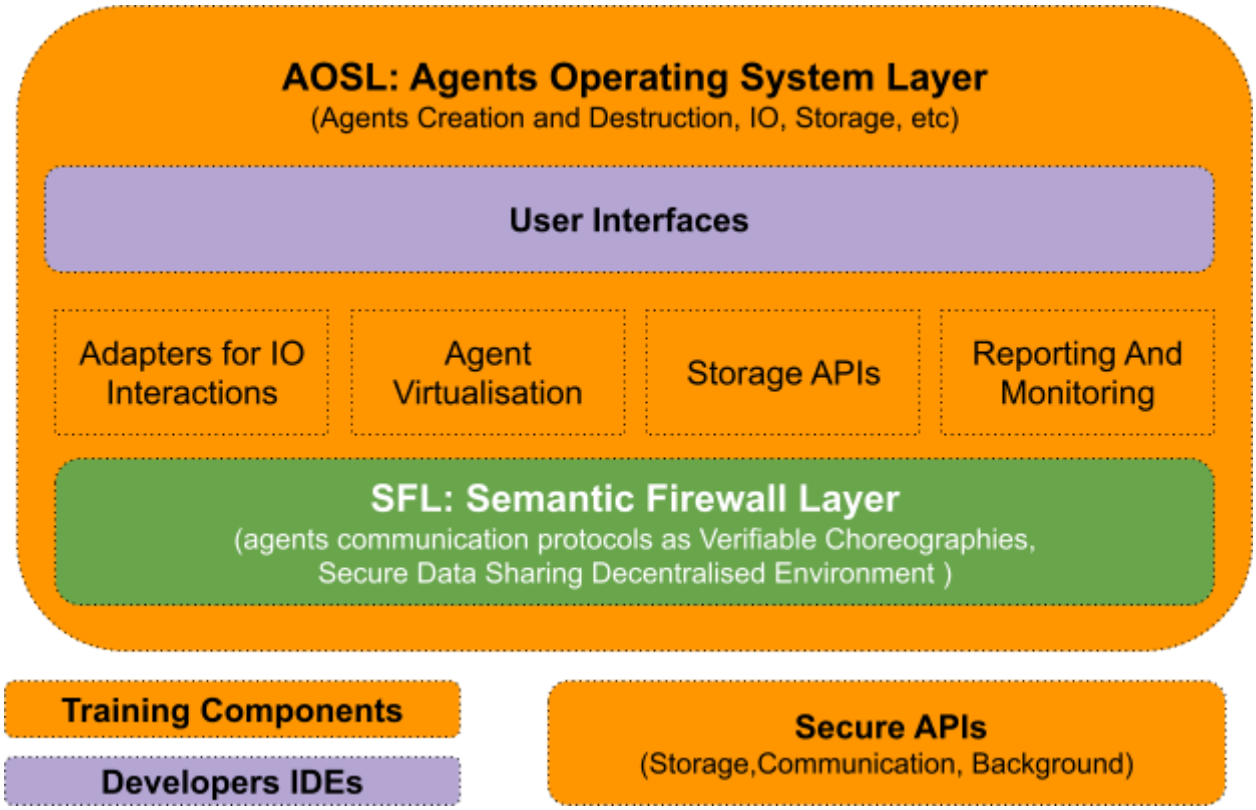
Future systems should consist of multiple agents with varied interests, memories, experiences, and sources of code and training data to guard against alignment threats that could arise accidentally or through deliberate manipulation. The most innovative concept we propose involves creating multi-agent 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. As we will detail in the following pages, our vision is to create an operating system where, instead of processes, there are agents with memory and persistence. These agents live only as long as they are needed to perform tasks. The more useful agents are rewarded by being brought back to life more frequently and by being able to create "copies," meaning they are cloned. Cloning can be useful both for verifying adherence to alignment and for perpetuating useful behaviours and experiences. The death of agents will also inevitably be necessary. The approach proposed in this report reminds us of the concept of reincarnation found in various religions, incorporating a cycle of life, death, and rebirth, adapted to the digital realm where agents evolve, replicate, and are selectively perpetuated based on their performance and alignment with desired goals.

Multi-Agent “Cognitive” Architecture for Guaranteed Alignment



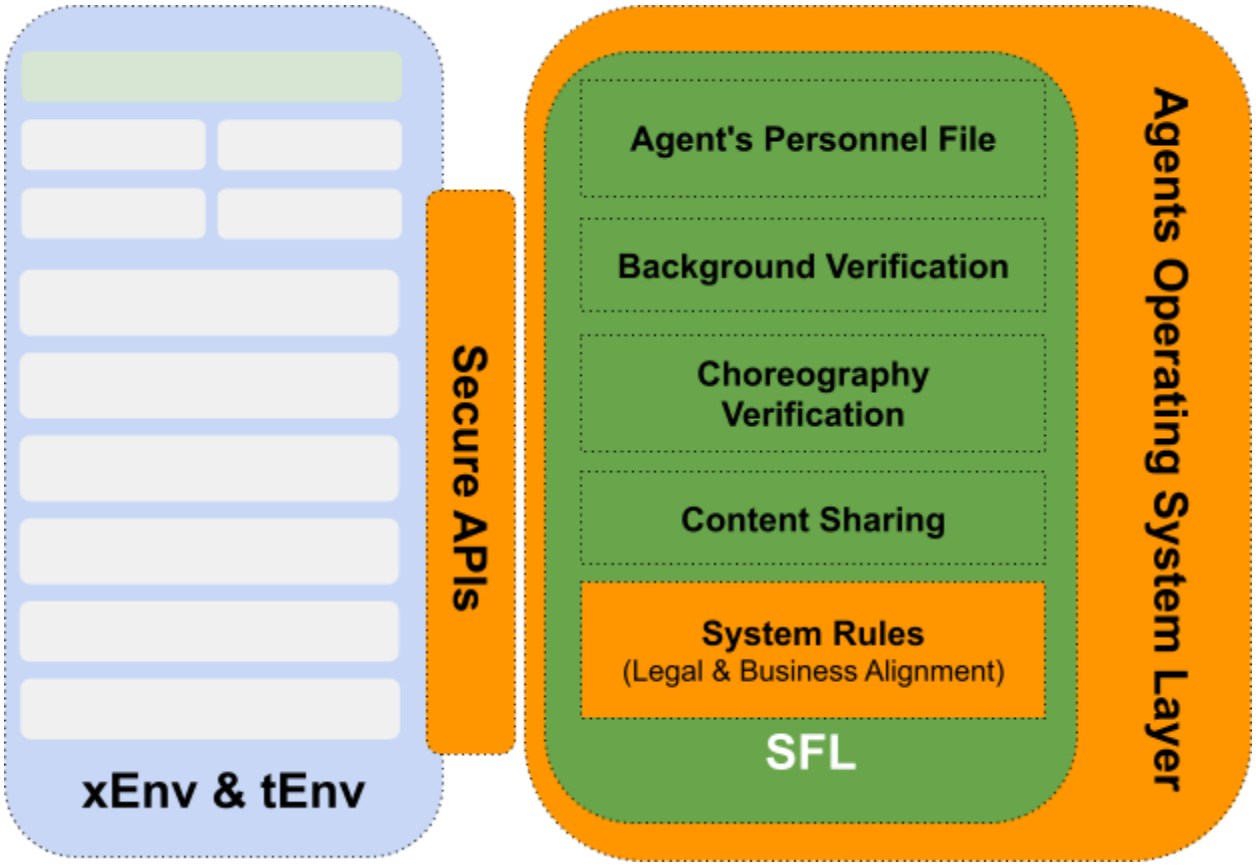
In our perspective, any multi-agent architecture designed to ensure the alignment of the entire swarm of agents with a set of legal and ethical values or rules will be structured around the following components: AOLS (Agents Operating System Layers), which provide hardware and software support for agents' instantiation and life cycles. This layer is not particularly intelligent; it is merely standard software code. Additionally, we have an Execution Environment (xEnv) that offers secure instantiation of agent instances, and a Training Environment (tEnv) that provides training tools and a playground for agents to test trained models or to create simulations of agents before deployment in production. Between all these layers, there is an SFL (Semantic Firewall Layer), which is a system that checks everything communicated between agents. The testing and execution environments cannot bypass the SFL. Various implementation modes of the SFL will be analyzed, but in principle, the SFL is a medium capable of analyzing the communication protocols between agents. The proposal is for these protocols to be implemented in the form of executable choreographies, more precisely, a form of verifiable choreography that is suitable for multi-agent environments.

Agents Operating System Layer (AOSL)



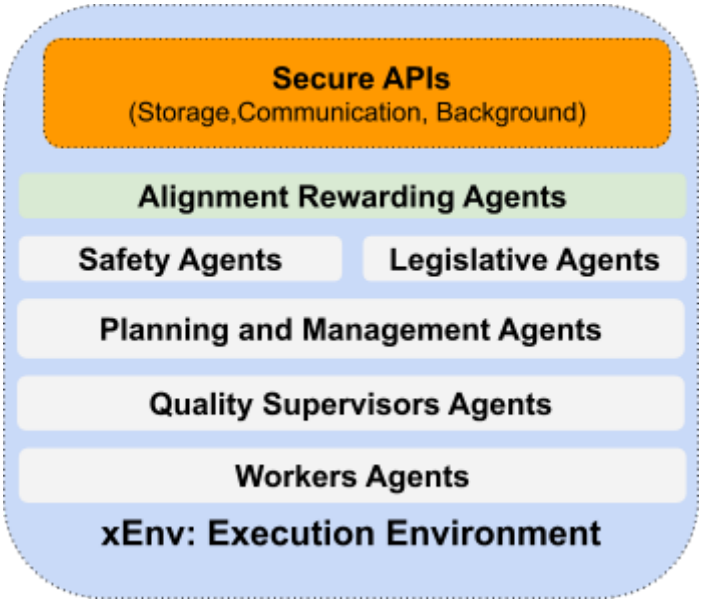
The AOSL (Agents Operating System Layer) is responsible for the creation and destruction of agents, I/O operations, and storage, but also for providing APIs for the creation of predefined tools and APIs that facilitate the creation of user interfaces for the system's users. This includes administrative aspects where we have APIs for reporting and monitoring, as well as for potential human users or clients served by the proposed architecture. For the creation and destruction of agents, we consider the necessity of APIs that can be grouped under Agent Virtualization APIs. These APIs ensure the operating system-level security of the agents' execution environment, preventing their communication via direct channels. Discovering hidden communication channels between malicious agents is the responsibility of the Semantic Firewall Layer, or as we saw in the first chapter, it could also be aided by the way executable choreographies are organized, which include agents specialized in detecting hidden channels or Advanced Persistent Threats (APTs). From this perspective, the Semantic Firewall Layer could be considered just a subcomponent of the operating system, but due to its significant importance, we will analyze it separately in a later chapter. It should also be mentioned that the code, APIs, tools, and user interfaces running virtually in secure (sandboxed) environments in tEnv and xEnv can be viewed as the user-space component in an operating system.

Semantic Firewall Layer



The SFL employs as much code as possible that is intelligible to human users or AI models with a high degree of explainability, avoiding the use of complex AI models that could conceal Advanced Persistent Threats (APTs), biases, and other attacks that could undermine alignment. Our research will focus on ways to accurately represent general rules related to alignment, as well as rules regarding compliance with laws and internal procedures proposed by the system owners. It is necessary for the rule description language to be as high-level as possible to allow for human auditing, yet sufficiently formal and comprehensive to encompass all types of nuances caught by verification components related to content sharing, choreography checking, and other background verifications executed by the SFL based on data collected about agents' behaviour and incidents. In this regard, the current vision of the SFL is to maintain information in an "Agent's Personnel File" and, upon detection of serious incidents or the accumulation of minor incidents that raise suspicions, to proceed with the destruction of the involved agent or agents, initiate agents with clean states, or even involve human oversight in the analysis of incidents.

Agents Execution Environment (xEnv)



The xEnv (Execution Environment) functions as a sophisticated virtualization space for multi-agent systems, resembling a Hypervisor in operating systems, offers APIs for secure and authenticated agent communication, enables secure data sharing among intelligent agents and supports background execution capabilities, alongside enforcing execution limitations based on the budget of each agent. Within xEnv, a diverse array of agents operates to maintain and optimize the system:

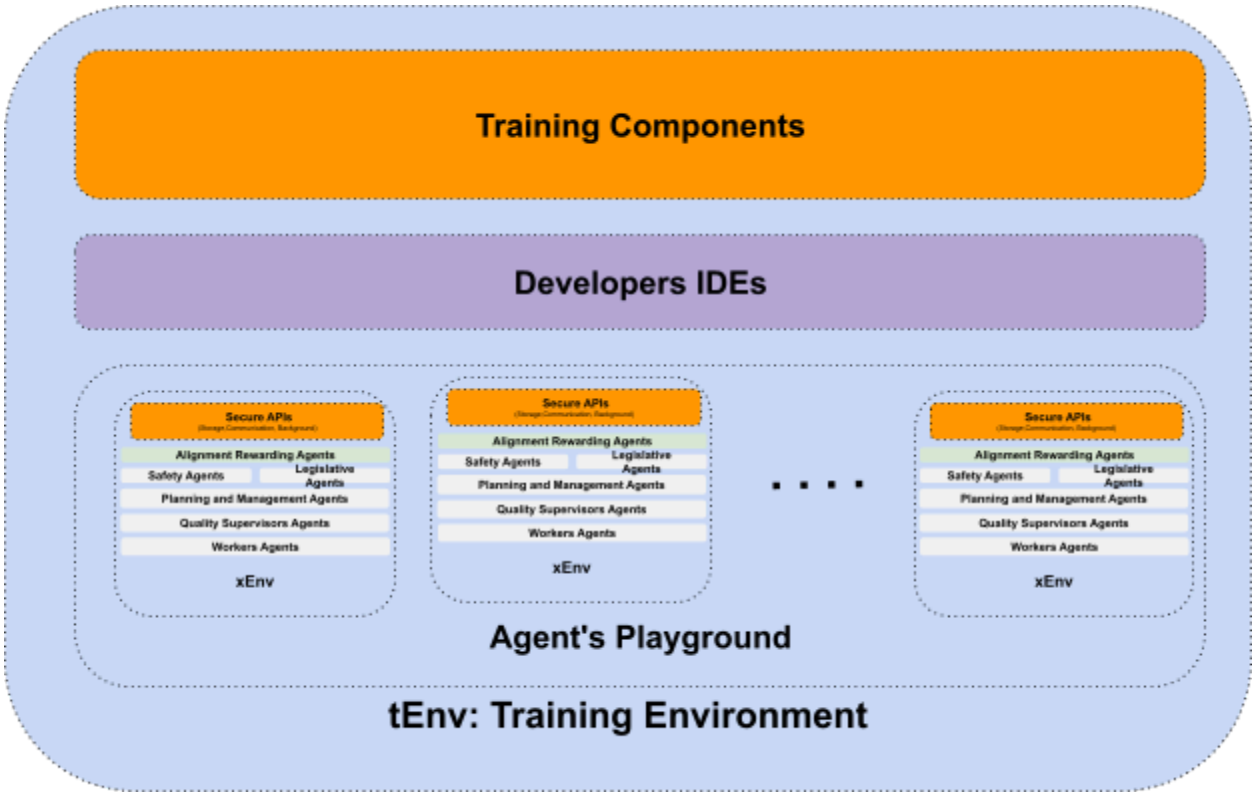
Alignment Rewarding Agents work in concert with the Semantic Firewall Layer (SFL) to identify and address any deviations from the system's alignment goals. Safety Agents serve as vigilant monitors over the activities of their peers, ensuring the environment remains secure and operations proceed without compromise. Legislative Agents are tasked with the creation and enforcement of rules, playing a crucial role in optimizing resource distribution and computational tasks across the environment. Planning and Management Agents take on roles akin to top or middle management in traditional organizational structures, tasked with orchestrating task execution. These agents can accumulate reputation or resources, which may justify their further instantiation. Alignment Rewarding Agents could have an influence within a xEnv regarding the resources allocated or the frequency of instantiation of various agents. This mechanism would allow for a dynamic adjustment of resources and instantiation frequency based on the alignment of agents with the desired objectives and behaviours. By rewarding agents that demonstrate alignment with predefined goals, the system can encourage the propagation of desirable behaviours and optimize the overall performance and efficiency of the multi-agent system. This approach ensures that agents contributing positively to the system's objectives are given precedence in resource allocation and opportunities for instantiation, thereby fostering an environment that promotes and maintains alignment with the system's core values and objectives.

Additionally, the environment hosts Quality Supervisors Agents and Worker Agents, specialized across dozens, hundreds, or thousands of different domains. The execution environment dynamically requests the instantiation of these agents based on the specific needs of concrete tasks. These agents not only have access to current task data but may also possess their own memory (or experiential knowledge) to optimize their performance.

This setup tries to ensure a robust, efficient, and secure operation of multi-agent systems, fostering an environment where agents can effectively collaborate, evolve, and contribute to the system's overarching goals while being on sleep, active or even destroyed accordingly with the need of the tasks and the safety and alignment concerns.

Somewhat surprisingly, this approach somewhat reminds us of the concept of reincarnation found in various religions. We mention this to reinforce the main intuition behind the proposed solution.

Training Environment (tEnv)



The tEnv (Training Environment) is envisioned as a comprehensive platform providing tools, libraries, and a user interface for developers to train AI models and experiment with swarms of agents before deploying them in actual production Execution Environments. The goal of a tEnv is to feature an Agent's playground designed to virtualize multiple Execution Environments in a cost-effective manner, implementing reduced budgets or other constraints to allow for deeper analyses and enhanced debugging capabilities. The Developer IDE are imagined as equipped with tools to examine data for biases, analyze the behaviour of the systems, and offer other tools for certifying the developed solutions, at least at the moment of deployment. This setup aims to ensure that developers have a robust environment for refining and validating their AI models, ensuring they meet the necessary standards and requirements before they go live, thereby enhancing the reliability and effectiveness of the deployed systems.

Why?

The proposed architecture introduces a layered design aimed at enhancing AI safety, ethical alignment, and operational efficiency. Central to this architecture is the capability for self-monitoring and self-regulation within AI agents, enabled by the Semantic Firewall Layer. This layer ensures AI operations adhere to ethical and legal standards, fostering societal trust through auditability. Enhanced transparency and accountability are crucial for meeting regulatory demands and building public confidence. Advanced security features within the architecture protect against external threats and internal errors, preserving the integrity of sensitive data. The decentralization aspect addresses vulnerabilities associated with centralized systems by introducing resilience and self-healing capabilities, further augmented by a unique "karma" system for agents that promotes positive behaviour over time. Interoperability and specialization across AI agents reduce complexity and enhance system manageability. The architecture's modular and scalable design ensures it can adapt to various scales of application, from small to large ecosystems. Lastly, its forward-looking design principles ensure adaptability to future technological developments, securing its long-term applicability in the dynamic field of AI. This comprehensive approach not only mitigates current AI alignment challenges but also paves the way for a more secure, ethical, and effective deployment of AI technologies.

In a landscape where superintelligence poses both immense potential and risks, the Semantic Firewall acts as a crucial safeguard. By design, it segments intelligence across multiple components, preventing any single part from becoming overwhelmingly powerful while still fostering collective growth. This segmentation ensures that while the system's overall capabilities can expand, they do so under strict ethical and operational guidelines. The firewall's role is to preemptively neutralize threats by enforcing these guidelines, making it exceedingly difficult for a superintelligence to bypass safeguards without detection. This architectural choice not only mitigates the risk of rogue AI behaviour but also aligns with the broader goal of developing AI that enhances societal well-being without compromising security or ethical standards.

The architecture adopts a zero-trust approach towards agents, actively resetting those veering towards potentially negative paths. Simultaneously, it aims to repurpose and amalgamate agents' experiences in a manner reminiscent of human societies and natural ecosystems. This strategy not only ensures ongoing vigilance against deviations but also fosters a collaborative, growth-oriented environment. By emulating societal and ecological dynamics, the architecture benefits from diversity and resilience, promoting a harmonious evolution of AI capabilities within a framework of mutual trust and ethical alignment.

Nature and science-inspired ideas requiring research

In the following table, we have summarized over 20 ideas that are inspired by nature and science, which require research to implement the various components described above, primarily focusing on the "Alignment Rewarding Agents" area. However, this research could also influence the Semantic Firewall Layer (SFL).

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	Psychology	Medium

	prioritization of information retention.		
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 sciences like biology, psychology, sociology or with philosophy through problems of epistemology, knowledge representations, ethics, and axiology.