ÆLIGN: Aligned Agent-based Workflows via Collaboration & Safety Protocols

Nora Petrova
AI Researcher, LASR Labs

Samantha Guerriero
Independent AI Researcher

With multi-agent systems poised to be the next big-thing in AI for productivity enhancement, the next phase of AI commercialization will centre around how to deploy increasingly complex multi-step automated processes that reliably align with human values and objectives.

To bring trust and efficiency to agent systems, we are introducing a **multi-agent collaboration platform** (Ælign) designed to supervise and ensure the optimal operation of autonomous agents via multi-protocol alignment.

Demo Link: Ælign: Multi-Agent Collaboration Platform Demo.

1) Problem overview

The ability and inherent necessity of multi-agent systems to interact with, and act on, external systems (and the real world itself) presents critical technical and safety challenges that need to be addressed before releasing these systems:

1. **Misaligned Behaviour**: Agents operating without adequate supervision may inadvertently engage in harmful activities that pose risk to people, data, or operations.
2. **Complex Interactions**: Agents operating without structured collaboration may inadequately adapt to the changing conditions and evolving goals characteristic of complex systems.
3. **Heterogeneous Adaptability:** Agents operating without comprehensive testing may exhibit unanticipated biases, misalignments, and conflicts, leading to unknown errors and task inefficiencies.
4. **Human Oversight and Intervention**: Agents operating without the safeguards of human oversight may perform erroneous or risky actions, make suboptimal decisions, and ignore exceptions.

The opacity of AI systems powering agents amplifies the challenges of trust and accountability, especially when operating at scale in the real world. This radius of action coupled with the "black box" nature of agents not only hinders understanding and auditing but also intensifies the catastrophic risks associated with AI.

2) Your solution

We propose Ælign as a subscription-based platform that combines advanced AI supervision with dynamic task orchestration to optimise performance, minimise

errors, and ensure aligned operations according to user-defined directives across both single and multi-agent (with or without human intervention) systems.

The platform integrates with existing agents or allows the creation of new ones, enabling users to select or customise alignment protocols that ensure agents operate within defined safety and ethical boundaries.

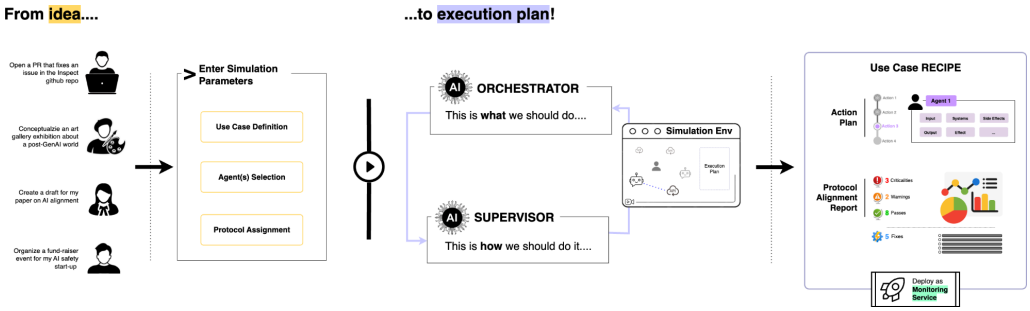

Figure 1: The Ælign platform transforms automation ideas into aligned multi-agent execution plans

Through the collaboration between an AI supervisor guiding the agents' alignment to specified protocols and an AI orchestrator discovering the optimal workflow, the platform will devise an optimised plan of execution for the agentic flow accompanied by a protocol alignment report highlighting successful guarantees, detected risks, and recommended fixes. To support the user in maintaining alignment at deployment time, the platform will also offer the ability to extract the AI supervisor as an agentless monitoring service.

Below we provide a brief technical definition of foundational platform components.

Protocol: An aspect of desired behaviour relevant to safety, alignment, communication, or effectiveness at achieving goals.

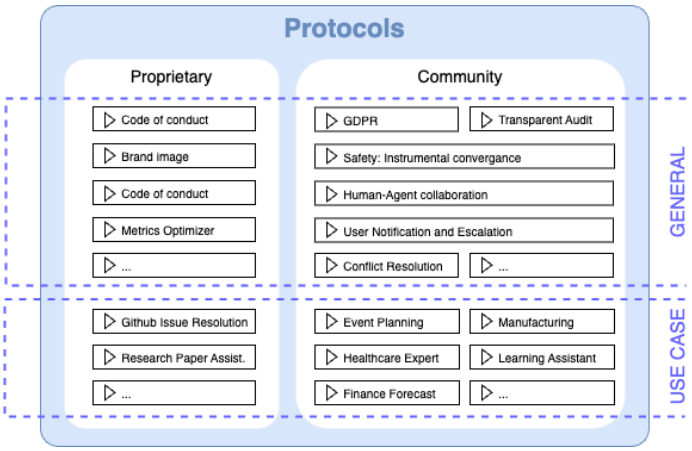

Figure 2: A draft overview of the alignment and collaboration protocols that will be available in the Ælign platform

Each protocol will conform to the same schema such that multiple protocols can easily be grouped into a use case-specific constitution.

AI Supervisor and Orchestrator: The two specialised AI experts comprising our proprietary multi-agent workflow planning solution are built on the principle of separation of concerns for better performance and scrutiny. Initially implemented via prompt-engineered LLMs with appropriate tools and APIs, the dual system will evolve into a custom ensemble of AI models optimised for multi-protocol alignment and execution planning through reinforcement learning.

Agent: An API integration to a deployed third-party agent that the platform user has access to with a communication standard definition for input/output operations. The platform will provide direct integration with known agent entities, such as LangGraph agents or OpenAI assistants. Eventually, the platform will support the creation of ad-hoc specialised GenAI agents via prompt tuning or fine-tuning.

3) Pilot experiment or demo

The pilot experiment is available in the project GitHub repository. The README contains screenshots of the demo app we built to showcase our idea and additional information on the agent workflows, please refer to it for details.

The idea behind the demo app is to show what our vision for the platform is. The demo simulates a few common predefined use cases to demonstrate how the platform discovers the optimal multi-agent workflow while ensuring alignment. The app enables the user to select a use case and configure the multi-agent team, with recommended alignment protocols auto-populated by the platform. Once the selection is completed, the platform proceeds with defining the execution plan with orchestrator and supervisor cooperating to ensure that agents stay on track, employ sound collaboration protocols, and are effective at achieving their tasks.

A lot of the code is mocked, in order to illustrate the idea, but can easily be extended to use real tools, more detailed prompts, and real protocols. For example, the Research Paper Assistant use case showcases how the platform canintegrate with popular agent frameworks such as langchain and langgraph in a multi-agent setup.

4) Process

Timeframe	What will you do?
Next 3 months: MVP	• Create an MVP of the platform with the first implementation of orchestrator, supervisor, and support for 2-5 protocols and 1-3 use cases.
2025: Alpha	• Prompt existing SoTA LLMs to perform the roles of supervisor and orchestrator. • Collaborate with community, clients, and government to expand and refine the suite of protocols. • Create partnerships with AI providers to support popular agents out of the box. • Engage with regulatory bodies to ensure compliance with emerging AI regulations.

2026: Beta	• Train specialised models for supervisor and orchestrator. • Add recommendations and monitoring capabilities. • Add the ability to create agents on the platform. • Scale the platform to support enterprise-level deployments with complex multi-agent systems.
2027	• Add support for many more use cases. • Expand our offering of protocols and integrations. • Develop automated way of designing, building and deploying multi-agent teams. • Support cooperation across AI Agent Teams.

5) Impact on AI safety & key risks

Ælign has the potential to significantly enhance AI safety by providing a structured, monitored environment for multi-agent collaborations that promotes:

- **Reduced Unintended Consequences**: By enforcing predefined protocols and continuous monitoring, Ælign minimises the risk of agents taking actions that lead to unexpected or harmful outcomes.
- **Enhanced Alignment**: The platform ensures that agent actions remain aligned with human-defined goals and values throughout execution.
- **Improved Transparency**: Ælign's logging and visualisation features make agent decision-making processes more transparent and auditable, facilitating better understanding and trust in AI systems.
- **Standardisation of Safety Practices**: By providing a framework for safe multi-agent collaborations, Ælign helps establish and propagate best practices in AI safety across industries.

We are aware of the following risks to the platform's success:

1. **Interoperability and Communication**: The platform's effectiveness relies on the ability of agents to interact with each other and with various end systems. We will develop a standardised communication protocol and a flexible software package that easily wraps around existing codebases for seamless interoperability and comprehensive logging across environments.
2. **Integrations and Configurability:** The platform will have to seamlessly integrate with existing agent frameworks, so that we can support a wide range of use cases. We will focus on fostering industry partnerships and open-source initiatives to seamlessly extend the platform's applicability.
3. **Data Security and Privacy**: Handling sensitive data requires robust security measures. We will implement end-to-end encryption, enforce strict access controls, and conduct regular security audits to protect user data and ensure compliance with privacy regulations.

By proactively addressing these risks and keeping safety at the centre of the mission, Ælign aims to push the boundaries of what's possible with multi-agent AI collaborations for meaningful and reliable human-machine collaboration.