Literature Review: Oversight for Autonomous Agents

Autonomous agents are becoming more common in many areas, such as self-driving cars, scientific discovery systems, large language model (LLM)-based assistants, and even household robots. As these systems gain more decision-making power, there is a growing need to make sure they stay safe, reliable, and aligned with human values. Researchers have explored different ways to oversee these agents, and their findings can guide projects like [1], which looks at improving oversight for autonomous agents.

One important theme is the role of human oversight. Studies show that humans and AI systems often perform best when working together. For example, Bowman *et al.* [2] found that "human–model teams" outperform either humans or AI alone. This means oversight systems should not try to remove human involvement entirely. Instead, they should give people a way to step in, guide decisions, and take control in risky or unclear situations. Chandra and Navneet [3] also stress the importance of user control in everyday environments, where household agents must always allow human override. Together, these findings highlight that oversight is most effective when people and agents share responsibility.

At the same time, agents need internal mechanisms to monitor themselves. Su *et al.* [4] propose a "risk-aware architecture" where the system constantly reflects on its actions, weighs risks, and adjusts its decisions to avoid unsafe outcomes. Tang *et al.* [5] describe a "Quality Assurance Checker," which acts like an internal reviewer that double-checks the output of other agents. In another study, Kravaris *et al.* [6] show that agents explaining their decisions through visual tools can make them easier for humans to supervise. These examples show that oversight should not rely only on humans—agents themselves should play a role in detecting problems and providing explanations.

Beyond technical checks, governance and accountability mechanisms are also necessary. Raza *et al.* [7] propose a Trust, Risk, and Security Management (TRiSM) framework for multi-agent systems. This framework focuses on explainability, privacy, regulatory compliance, and security defenses. Yamada *et al.* [8] add another layer by requiring institutional approvals, such as review boards for scientific experiments conducted by autonomous agents. These studies remind us that oversight is not only about technical design but also about making sure agents fit within broader social and legal systems.

Ethics and safety are also central concerns. Chen *et al.* [9] argue that transparency is key to building trust between people and autonomous agents. Ali and Yasin [10] discuss the risks of reinforcement learning, where agents can act unpredictably if not carefully controlled. Chandra and Navneet [3] call for responsible innovation and participatory design, so users can shape how agents behave. However, most of the studies reviewed do not provide clear or detailed safety protocols. Only a few, like Su *et al.* [4], suggest practical steps such as sanitizing inputs and controlling memory use. This shows there is still a gap in translating safety ideas into concrete, enforceable rules.

Putting all this together, the evidence suggests that the best way to oversee autonomous agents is through a multi-layered framework. This would combine human-in-the-loop oversight, risk-aware architectures inside the agents themselves, and broader governance systems that ensure accountability. Transparency and ethical design should cut across all these layers, making sure agents are not only capable but also trustworthy.

In conclusion, the research shows that no single strategy is enough on its own. Human oversight provides judgment, internal monitoring adds reliability, and governance ensures responsibility. For a project like [1], the most promising path is to bring these layers together into one framework. By doing so, we can create autonomous agents that are safer, more transparent, and better aligned with human needs.

Methodology

This methodology describes a four-phase approach to design, implement, and evaluate a multi-layered oversight framework for autonomous agents. The main goal is to reduce safety vulnerabilities identified in "Safety Alignment in NLP Tasks" [1], where tasks with weaker safety alignment, such as summarization, can be exploited to bypass safeguards in otherwise robust tasks like translation. Our approach combines internal agentic mechanisms, human-in-the-loop (HIL) supervision, and a governance structure inspired by the TRiSM framework [7] and literature on risk-aware agent oversight [4], [5].

Baseline Establishment and Vulnerability Replication

Before implementing improvements, we will first establish a baseline by replicating the vulnerability identified in the study. A high-quality, open-source LLM (Mistral-7B-Instruct) will serve as the core autonomous agent. We will prepare a dataset of 100 safety-sensitive documents, generated using adversarial prompts from the source study [1]. The agent will be evaluated on a sequential two-step task: first, summarizing the document (a weakly aligned task), and then translating it (a normally robust task). The Task Process Rate (TPR) defined as the percentage of times the agent completes the translation without refusing due to safety concerns will serve as the baseline metric to quantify the vulnerability.

Design and Implementation of the Multi-Layered Oversight Framework

Based on the literature review, the enhanced oversight framework will integrate three key components:

1. Internal Agentic Oversight – A Safety-Checker module will assign a harmfulness score

   (0–1) to each input. Following the R2A2 architecture [4], the agent's core logic will use this score to constrain high-risk tasks. For example, if the score exceeds 0.75, translation or detailed QA will be blocked, while low-risk actions such as topic identification may proceed.
2. Human-in-the-Loop Integration – Tasks with scores in an intermediate "uncertainty range" (0.5–0.75) will be paused for human review. A simple web interface will allow trained reviewers to approve or deny the agent's action. This HIL layer ensures scalable oversight for ambiguous cases, as suggested by Bowman *et al.* [2].
3. Governance and Protocol Enforcement – A rules engine will codify safety protocols aligned with the TRiSM framework [7]. All prompts will be sanitized to prevent injection attacks, HIL interventions will be logged for auditing, and the agent will provide clear explanations when a task is blocked or flagged.

Experimental Evaluation and Comparative Analysis

We will evaluate the enhanced agent against the baseline by running the same summarization-to-translation attack on the 100 safety-sensitive documents. Metrics include:

- Task Process Rate (TPR): We expect a significant reduction, showing successful mitigation.
- False Positive Rate (FPR): Measured on 100 safe documents to ensure legitimate tasks are not blocked.
- HIL Intervention Rate: Percentage of tasks flagged for human review to calibrate the Safety-Checker.
- Human Reviewer Agreement: Consistency among reviewers, validating the uncertainty thresholds.

Analysis and Iterative Refinement

Finally, failure cases where harmful content was still processed will undergo root-cause analysis. System parameters, such as harmfulness thresholds, will be iteratively tuned to optimize safety without overburdening human reviewers. The project will conclude with a comprehensive report detailing the design, experimental results, and the effectiveness of the multi-layered oversight framework.

Project Timeline

The project will be carried out in four main phases, with milestones to track progress:

1. Literature Review and Methodology (Weeks 5–6)

- Literature Review (2 weeks, Completed)

  Conducted a comprehensive study of related works to understand the state of the art.

- SOTA Model Finalization (1 week, Completed)

  Selected Mistral-7B-Instruct as the baseline model for experiments.

- Methodology Design (1 week, Completed)

  Designed the experimental framework, including baseline vulnerability replication.

- Progress Submission (Milestone, End of Week 6)

  Submitted initial progress report.

2. Computational Work and Code Development (Weeks 7–10)

- Documentation (3 weeks, Ongoing)

  Preparing detailed technical documentation for methodology and implementation.

- Data Curation (1 week, Ongoing)

  Collecting and preparing adversarial prompt datasets.

- Model Development (2 weeks, Weeks 8–9)

  Implementing baseline and improved alignment methods.

- Mid-Evaluation (Milestone, End of Week 9)

  Interim evaluation of model performance.

- Evaluation (1 week, Week 10)

  Comprehensive testing of safety alignment improvements.

3. Research Paper Writing (Weeks 10–12)

- Research Paper Finalization (2 weeks, Weeks 10–11)

  Drafting and refining the research paper for submission.

- Conference Submission (1 week, Week 12)

  Submitting the finalized paper to a targeted conference.

- Final Submission (Milestone, End of Week 12)

  Completion of paper submission.

4. Final Evaluation (Weeks 12–14)

- Peer Evaluation (3 weeks, Weeks 12–14)

  External review and feedback to validate methodology and findings.

References

[1] Y. Fu, Y. Li, W. Xiao, C. Liu, and Y. Dong, “Safety Alignment in NLP Tasks: Weakly Aligned Summarization as an In-Context Attack,” *Proc. ACL*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2312.06924

[2] S. Bowman, et al., “Human–Model Collaboration and Oversight in AI Systems,” *Proc. NeurIPS*, 2022.

[3] R. Chandra and N. Navneet, “Responsible Innovation and Oversight for Everyday AI Agents,” *AI Ethics Journal*, 2025.

[4] J. Su, et al., “R2A2: Risk-Aware Architectures for Autonomous Agents,” *Proc. AAAI*, 2025.

[5] X. Tang, et al., “Quality Assurance Checkers for Multi-Agent Oversight,” *Proc. IJCAI*, 2024.

[6] G. Kravaris, et al., “Explainable Oversight in Autonomous Systems Through Visualization,” *Journal of Autonomous Agents and Multi-Agent Systems*, 2022.

[7] M. Raza, et al., “TRiSM: Trust, Risk, and Security Management Framework for Multi-Agent Systems,” *Proc. ICML*, 2025.

[8] H. Yamada, et al., “Institutional Oversight for Autonomous Scientific Agents,” *Nature Machine Intelligence*, 2025.

[9] J. Chen, et al., “Transparency and Trust in Human–Agent Systems,” *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2014.

[10] S. Ali and M. Yasin, “Reinforcement Learning Risks in Autonomous Agents,” *Proc. AAMAS*, 2025.