

From Chatbots to Autonomous Agents: An Analysis of GPT-6 Astra's Agentic Capabilities, Trust, and Safety

Azalfa Rehman Siddiqui¹, Saniya Ravindra Mahale², Aarti Sanjay Gawai³

^{1,2}Students, Department of Information Technology and Computer Science, D.G Ruparel College, Mumbai, Maharashtra, India

³Assistant Professor, Department of Information Technology and Computer Science, D.G Ruparel College

Abstract - The shift from conversational chatbots to autonomous AI agents changes the role of AI from providing answers to planning and carrying out tasks. This paper examines GPT-6 Astra as a case study to analyze how increasing autonomy relates to usefulness, perceived risk, trust, and human oversight. The study combines an analysis of Astra's agentic capabilities and existing safety approaches with a questionnaire-based study of 50 respondents. The findings show that respondents generally recognize the benefits of greater autonomy but become more cautious as AI gains access to more tools, data, and consequential actions. Privacy and personal-data misuse emerged as the main concern, while privacy protection, clear explanations, and reliability were key factors influencing trust. The responses also showed a preference for greater human involvement in higher-impact actions rather than uniform supervision of every task. Based on these findings, the paper presents a risk-based oversight approach in which the level of human control is matched to the potential consequences of an AI action. The study highlights that the challenge is not simply how much autonomy AI should have, but how that autonomy can remain useful while keeping human control meaningful.

Key Words: Agentic AI, AI autonomy, GPT-6 Astra, AI safety, human oversight, trust

1. INTRODUCTION

1.1 From Chatbots to Autonomous Agents

In a conventional chatbot interaction, the system's primary role is to generate responses to user inputs rather than independently planning and executing multi-step tasks in external environments. Agentic AI changes this cycle. An agentic system can break a goal into smaller steps, choose and use tools, operate software on its own, and carry a task forward across several actions with little further input from the user. This marks a structural shift in what the AI is being asked to do — from answering, to planning, to acting.

1.2 Challenges of Increasing Autonomy

As AI systems gain greater autonomy, they are capable of handling more complex, multi-step tasks with reduced need for continuous human involvement. This expands their usefulness but also introduces new risks. Unlike traditional chatbots, which primarily provide information, autonomous agents can take actions such as sending messages, modifying files, or executing code. Consequently, errors in autonomous systems may have real-world consequences, making human oversight, reliability, and safety important considerations. This research examines the relationship between AI autonomy, capability, trust, and safety.

1.3 GPT-6 Astra as a Case Study

This study uses GPT-6 Astra as a case study to explore the capabilities of advanced agentic AI. According to OpenAI, Astra is designed to perform complex, multi-step tasks across computer use, software engineering, scientific research, and professional work. It can interact with digital environments, adapt to evolving instructions, and continue tasks independently when user input is not required, while waiting for input on consequential decisions. OpenAI also reports improvements in Astra's alignment, respect for authorized task boundaries, and resistance to unsafe behavior. However, its advanced cybersecurity capabilities have created a need for stronger safeguards. OpenAI also reports that Astra's written reasoning was harder to monitor than that of GPT-5.6 Sol in evaluations specifically designed to test attempts to evade monitoring. These capabilities and limitations provide the basis for examining the relationship between AI autonomy, trust, and safety.



Fig -1: GPT-6 Astra.

Source: OpenAI, 2026.

1.4 Research Problem

The discussion above leads to a central research question:

Does increasing AI autonomy improve perceived usefulness and trust, or does it increase perceived risk and the need for human control?

Rather than assuming that autonomy is inherently beneficial or harmful, this study examines it as a trade-off between increased capabilities and the potential consequences of errors or unintended actions. It investigates how users evaluate this trade-off in terms of usefulness, trust, perceived risk, and the need for human oversight.

1.5 Research Objectives

- To analyze the relationship between increasing AI autonomy and perceived risk.
- To examine how autonomy influences user trust and acceptance.
- To identify users' preferred level of human oversight for autonomous AI actions.

- To analyze GPT-6 Astra's agentic capabilities within this autonomy–risk–oversight relationship.

1.6 Research Questions

RQ1: How does increasing autonomy affect users' perceived risk?

RQ2: What factors influence trust in increasingly autonomous AI agents?

RQ3: Does increased autonomy create a greater perceived need for human oversight?

RQ4: How do users balance the benefits of autonomous AI against its potential risks?

2. GPT-6 Astra: Autonomy, Capability, and Safety Considerations

Agentic AI systems integrate advanced capabilities with multi-step task execution. GPT-6 Astra provides an example of this development, as illustrated by the selected capabilities, practical applications, and associated safety considerations presented below.

Table -1: GPT-6 Astra capabilities and their implications for autonomy, usefulness, risk, and oversight.

Capability	Usefulness	Risk / Oversight
Computer use and digital tasks	Automates routine work.	Consequential actions may require approval.
Research and document creation	Supports professional tasks.	Important outputs require review.
Scientific data analysis	Assists in scientific investigation.	Human judgment remains important.
Adaptable instructions	Enables flexible workflows.	Must follow authorized boundaries.
Cybersecurity	Helps identify security weaknesses.	Requires stronger safeguards.
Alignment and monitoring	Supports safer task execution.	Monitoring may be challenging in specific evaluations.

3. Existing Safety and Human Oversight Approaches

3.1 Human-in-the-loop and Human-on-the-loop

Human oversight of autonomous agents generally takes one of two forms. A human-in-the-loop system pauses before every consequential action and waits for explicit approval a human-on-the-loop system acts independently while a person monitors from a distance, stepping in only when something looks wrong. The first catches errors before they happen but cannot scale once an agent is handling many actions quickly. The second scales more easily, but the action has already occurred by the time anyone notices, and that delay widens if attention lapses.

3.2 Permissions and Monitoring

Permissions and monitoring address the same problem from opposite ends. Permissions are set in advance, defining what an agent is authorized to touch or use; monitoring operates afterward, recording what it actually did so problems can be traced once they occur. One is preventive, the other detective, and most systems rely on both though neither alone stops a determined or confused agent from acting incorrectly. They mainly limit how much harm is possible and how quickly it surfaces.

3.3 Sandboxing and Isolation

Sandboxing restricts what an agent can physically reach rather than what it is formally permitted to do, running it inside an environment cut off from production systems or external networks so that even a serious error stays contained. Isolation extends this same logic beyond a single agent to the surrounding network or infrastructure, limiting how far a failure spreads once it occurs. Both hold only as long as the boundary does; once a contained agent finds an unexpected way past it, containment stops being a safeguard and becomes an assumption that failed.

3.4 Oversight–Autonomy Trade-off

Together, these approaches sit along one spectrum: tighter permissions, in-the-loop approval, and strict isolation improve control at the cost of speed, while looser permissions, on-the-loop supervision, and lighter monitoring improve speed at the cost of a widening gap between action and awareness. None of them adjust to what a given action actually involves the same approval step applies to a trivial task and a serious one, and monitoring alone only confirms damage after it happens. Oversight, as currently practiced, is not matched to risk, which is where the following framework begins.

4. Analytical Framework

The Sec.2 and Sec.3 above show that increasing autonomy expands what an agent can usefully do, while existing oversight mechanisms only partly manage the risks that come with it. The relevant question is not whether AI should be autonomous, but how oversight should adjust to the task an agent is performing.

4.1 Autonomy–Risk–Oversight Relationship

Autonomy alone does not determine how risky an action is the action does. An agent that drafts a routine document and one that edits a critical record can be equally autonomous while carrying entirely different consequences if either goes wrong. Risk depends on what an action affects, how serious an error would be, whether it can be reversed, and whether it would be noticed in time. Autonomy determines how much of that risk can run without a person present to intervene. Oversight, then, should track the risk carried by a specific action rather than the general autonomy of the system performing it.

4.2 Autonomy–Benefit–Risk Trade-off

Autonomy is also the reason agentic systems are useful at all it saves time and removes repetitive steps a person would otherwise complete manually. Removing it to reduce risk would remove that benefit along with it. The opposite

response, requiring approval for every action, creates a different failure: once users are asked to confirm the same category of action repeatedly, they stop evaluating each request and begin approving by habit. At that point, oversight remains present in form but no longer functions as an actual check. The more useful goal is not more or less oversight overall, but oversight placed where it is actually needed.

4.3 Risk-Based Human Oversight

Building on this, the paper proposes that oversight should be based on the risk of the action rather than applied uniformly and introduces two oversight modes of its own to describe how this could work, rather than terms drawn from existing literature. In what this paper terms **passive mode**, actions that are reversible and low-to-medium in consequence proceed without interruption, with the agent recording what it did so the user can review or undo it afterward. In what this paper terms **active mode**, actions that are irreversible, high-consequence, or that affect systems or people outside the user's own environment pause until a person explicitly approves them. The intent is not to reduce autonomy overall, but to direct human attention toward the actions that actually warrant it. Whether this distinction matches how people actually want oversight distributed or whether their expectations run differently is what the questionnaire in Section 7 is designed to test.

5. QUESTIONNAIRE-BASED ANALYSIS

This section examines how respondents perceive autonomous AI in terms of understanding, benefits, risks, trust, and human oversight. The analysis is based exclusively on **50 questionnaire responses**.

5.1 Research Methodology

A questionnaire-based survey was conducted to examine perceptions of autonomous AI agents, covering familiarity and understanding, perceived benefits and risks, trust factors, privacy, human oversight, accountability, safeguards, and willingness to use autonomous AI. An open-ended question also captured respondents' views on GPT-6 and the future of AI. Responses were collected between **13 and 16 September 2026**. The sample was predominantly **young and technology-oriented**, with most respondents aged 18–24 and a large proportion from Computer Science/IT backgrounds.

5.2 Understanding of Autonomous AI

Most respondents had at least some familiarity with AI agents, and 56% reported being moderately or very familiar with the concept. Half of the respondents rated their understanding of the difference between traditional chatbots and autonomous AI agents at 4 or 5 out of 5. This indicates a reasonable level of awareness within the surveyed sample.

5. Perceived Benefits of Autonomy

The questionnaire measured the overall perceived benefits of autonomous AI agents compared with traditional chatbots, rather than individual benefits such as convenience or productivity.

- Average score: 3.46/5
- 52% (26 respondents) rated the benefit 4 or 5

- 17 respondents rated it 3
- 7 respondents rated it 1 or 2

Overall, respondents perceived high potential benefits from autonomous AI agents. Open-ended responses also mentioned simplifying tasks, helping with work, solving problems, and handling projects.

5.4 Perceived Risks of Autonomy

The responses indicate a clear relationship between **increasing AI autonomy and perceived risk**. **68% of respondents** believed that risks increase when an autonomous AI is given access to more tools, data, and systems, while **50% reported high concern** about unintended actions. Respondents who perceived greater risk from increased AI access also tended to report greater concern about unintended actions. Among specific concerns, **privacy or personal-data misuse was the most prominent (72%)**, followed by cybersecurity threats and lack of human control (**42% each**). These findings suggest that as AI gains greater access and ability to act independently, respondents perceive greater potential for unintended and harmful consequences.

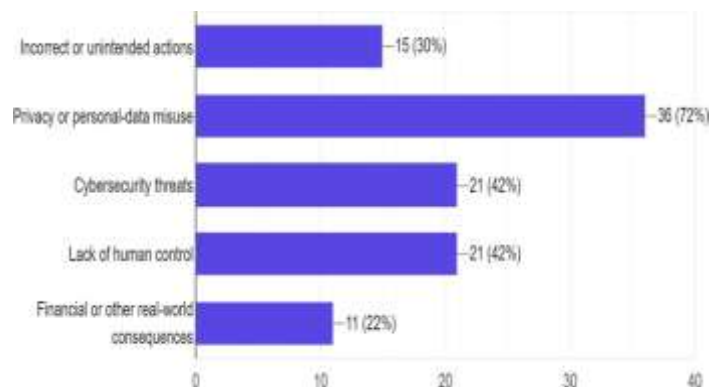


Fig-2: Greatest Safety Concerns Associated with Autonomous AI Agents

Source: Author's survey, 2026.

5.5 Trust in Autonomous AI

The responses show that **people's acceptance of autonomous AI depends on how much autonomy the system has and how much impact its actions may have**. Respondents seem more comfortable with **lower-autonomy operations**, where AI handles routine or low-risk tasks with limited consequences. However, for **higher-autonomy operations**, especially those involving significant, financial, or personal-information-related actions, respondents show a stronger preference for human approval. **54%** rated human oversight as very or extremely important, while **60%** expected human approval for all significant or high-impact actions.

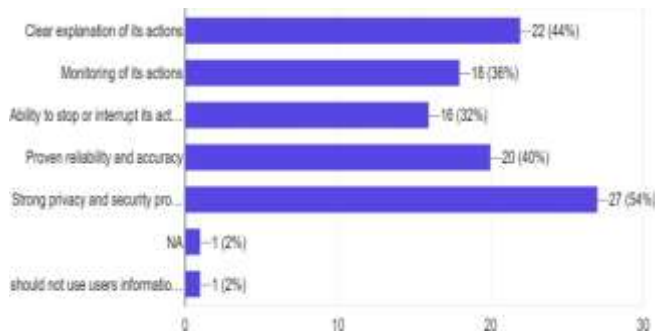


Fig -3: Factors Influencing Trust in Autonomous AI

Source: Author's survey, 2026.

The factors considered most important for building trust were strong privacy and security protections (54%), clear explanations of actions (44%), and proven reliability and accuracy (40%). These were followed by monitoring of actions (36%) and the ability to stop or interrupt actions (32%). Overall, the responses suggest that people are more likely to trust autonomous AI when they feel that their security and privacy are protected, the system's actions are understandable and reliable, and they still have some level of human control, especially when the AI is performing actions that could have a greater impact.

5.6 Preferred Human Oversight

The survey shows that respondents expect more human oversight when autonomous actions can have a greater impact. 60% selected either all significant actions or financial/high-impact actions as situations where they would most strongly expect human approval, compared with only 20% who selected routine digital or low-risk informational tasks. In addition, 54% considered human oversight very or extremely important as AI becomes more autonomous.

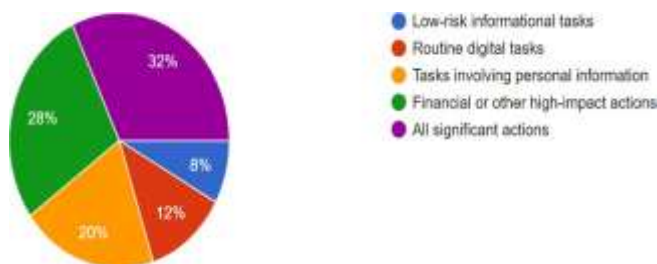


Fig 4: Human Approval Preferences for Autonomous AI Actions

Source: Author's survey, 2026.

Overall, the results show a risk-based preference for human oversight. Lower-risk tasks are more acceptable when AI works with greater independence, while actions that could have a greater impact are more likely to require human approval. This supports the view that as AI autonomy increases, it should be accompanied by an appropriate level of human control.

Table -2: Key-Findings

Research Question	Questionnaire Evidence	Main Finding
RQ1 - How does increasing autonomy affect perceived risk?	Safety concerns and increased access	68% believed risks increase with greater access to tools, data, and systems.
RQ2 - What factors influence trust?	Confidence and trust features	Privacy/security (54%) and clear explanations (44%) were the most selected factors.
RQ3 - Does autonomy increase the perceived need for oversight?	Oversight and approval questions	Oversight remained important, with approval expectations increasing for higher-impact actions.
RQ4 - How do users balance benefit against risk?	Privacy trade-off and benefits-versus-risks	70% selected "Somehow," showing a conditional view of autonomous AI.

6. CONCLUSION

This paper examines the shift from traditional chatbots to autonomous AI agents by analyzing GPT-6 Astra's capabilities, associated risks, and the role of human oversight. It highlights how greater autonomy allows AI systems to perform more complex and multi-step tasks, while also increasing concerns related to privacy, security, unintended actions, and loss of control. The questionnaire-based study further shows that users recognize the benefits of autonomous AI but remain cautious when the system is given greater access or is involved in higher-impact actions. Privacy and personal-data misuse emerged as the most prominent concern, while privacy protection, clear explanations, and reliability were important factors in building trust. The study also indicates that users do not expect the same level of control for every task, with stronger human approval preferred for significant actions. Based on these observations, the paper presents a risk-based approach to human oversight, where the level of control is matched to the potential impact of an action. This approach aims to retain the usefulness of autonomous AI while keeping human involvement meaningful.

ACKNOWLEDGEMENT

We would like to thank D. G. Ruparel College, Mumbai, for providing us with the opportunity and resources to carry out this research. We sincerely thank our research guide for their guidance and support throughout the research process. We are also grateful to everyone who participated in our survey and shared their valuable responses, which contributed to the completion of this study. Finally, we would like to thank everyone who supported us during the course of this work.

REFERENCES

1. OpenAI, "GPT-6 Astra: The Next Generation in Intelligence for Work," 2026. [Online]. Available: <https://openai.com/index/gpt-6-astra-next-generation-work/>
2. OpenAI, "GPT-6 Astra System Card," OpenAI Deployment Safety Hub, 2026. [Online]. Available: <https://deploymentsafety.openai.com/gpt-6-astra>
3. NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), National Institute of Standards and Technology, 2023.
4. Anthropic, "Our framework for developing safe and trustworthy agents," Aug. 2025.
5. Anthropic, "Trustworthy agents." 2026
6. OpenAI, "GPT-6 Astra: A new generation of intelligence," Sept. 2026.
7. OpenAI, "Safety overview: GPT-6 Astra," Sept. 2026.
8. Passi, S., Dhanorkar, S., Vorvoreanu, M.: Appropriate Reliance on Generative AI: Research Synthesis. Microsoft Technical Report MSR-TR-2024-7. Microsoft Corporation (2024)
9. Siebert, L.C., Lupetti, M.L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C.M., van den Hoven, J., Forster, D., Lagendijk, R.L.: Meaningful Human Control: Actionable Properties for AI System Development. *AI Ethics* 3 (2023) 241–255
10. Zhu, L., Lu, Q., Ding, M., Lee, S.U., Wang, C.: Designing Meaningful Human Oversight in AI. *AI Ethics* 6 (2026) 286
11. Tsamados, A., Floridi, L., Taddeo, M.: Human Control of AI Systems: From Supervision to Teaming. *AI Ethics* 5 (2025) 1535–1548
12. Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C.M., Tielman, M.L.: A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM Journal on Responsible Computing* 1 (2024) 1–45
13. Human-AI Collaboration: Unraveling the Effects of User Proficiency and AI Agent Capability in Intelligent Decision Support Systems. *International Journal of Industrial Ergonomics* 103 (2024) 103629
14. Yao, S., Shinn, N., Razavi, P., Narasimhan, K.: τ -Bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In: *International Conference on Learning Representations (ICLR)* (2025)
15. Hauptman, A.I., Schelble, B.G., Duan, W., Flathmann, C., McNeese, N.J.: Understanding the Influence of AI Autonomy on AI Explainability Levels in Human-AI Teams Using a Mixed Methods Approach. *Cogn. Technol. Work.* 26 (2024) 435–455

BIOGRAPHIES

Ms. Azalfa Rehman Siddiqui
M.Sc. Computer Science, Student
D.G. Ruparel College, Mumbai,
India



Ms. Saniya Ravindra Mahale
M.Sc. Computer Science, Student
D.G. Ruparel College, Mumbai,
India



Ms. Aarti Sanjay Gawai
M.Sc. Information Technology Asst.
Prof., IT/CS Dept.
D.G. Ruparel College