
AI Governance for Medical Chatbots: Designing a Multi-Agent Controller (MAC) for Safety¹

Quentin Marquet
42 School

Shakthivel Murugavel
42 School

Xavier Charles
42 School

Ouafae Moudni
42 School

Elise Racine
University of Oxford

With

In collaboration with Apart Research

Abstract

Unregulated medical AI chatbots pose significant risks, from dispensing incorrect health advice to exposing sensitive data. This research addresses the problem by proposing a Medical Agent Controller (MAC)—a multi-agent governance framework that audits and filters a chatbot’s responses in real time. We draw motivation from the growing adoption of AI-powered healthcare assistants amid mounting concerns over hallucinations, algorithmic bias, adversarial exploits, privacy risks, and the broader societal impact of AI cognition and machine psychology (Uppsala University, 2024; Blease et al., 2024; Zakka et al., 2023; Templin et al., 2024; Chang et al., 2024; Haque et al., 2025; Lee, Bubeck, and Petro, 2023; Omiye et al., 2023; Obermeyer et al., Weidinger et al., 2021; Zou et al., 2023; Ji et al., 2022; Alowais et al., 2023). Our approach leverages CrewAI, an open-source multi-agent platform, to orchestrate specialized agents—a verification agent and a red-teaming agent—alongside retrieval-augmented generation (RAG) for medical fact-checking. By coordinating these agents, MAC integrates behavioral analysis and human-AI interaction principles to mitigate risks, ensuring medical AI aligns with evolving social norms and institutional transformation in healthcare governance. A prototype implementation demonstrates that this approach can successfully identify and correct problematic medical advice, as detailed in our evaluation. We anticipate that MAC will enhance the reliability of medical chatbots and reduce the likelihood of harmful outcomes. In turn, this work highlights a potential AI governance mechanism for high-stakes applications, emphasizing the need for robust oversight in human-AI systems and regulatory compliance in healthcare deployments.

¹ Research conducted at the Women in AI Safety Hackathon, 2025

Keywords: machine psychology, human-AI interaction, institutional transformation, social norms, AI governance, risk mitigation, medical AI, healthcare chatbots, multi-agent systems, adversarial testing, red team simulation

1. Introduction

1.1 Problem Statement

AI-driven medical chatbots are increasingly used by both patients and clinicians for health information and decision support. However, without proper oversight or regulation, these chatbots can produce harmful misinformation or unsafe recommendations, leading to real-world risks. Notable failures include Babylon Health’s chatbot, which provided dangerously incorrect medical advice, and the Tessa chatbot, which—despite being designed to prevent eating disorders—ended up dispensing weight-loss tips that could be harmful to vulnerable users (Heather, 2018; Das, 2019; Wells, 2023). Unlike regulated medical devices or professional advice, most general-purpose chatbots (e.g., GPT-based assistants) operate in a gray area of oversight (Lomas, 2021). This raises the question: How can we ensure the safety and trustworthiness of medical AI chatbots in the absence of formal regulation?

The core problem addressed in this work is mitigating the risks associated with unregulated medical AI chatbots. Our goal is to develop a governance framework that can catch and correct harmful outputs in real time, thereby advancing AI safety in the healthcare domain.

1.2 Background and Motivation

AI-powered healthcare chatbots have surged in popularity, touted as a solution to resource constraints in medicine (Altamimi, 2023). Recent surveys indicate that nearly 20% of physicians have integrated tools like ChatGPT into their clinical workflows (Uppsala University, 2024; Blease et al., 2024; B Altamimi, 2023). While these bots assist with tasks ranging from drafting documentation to suggesting preliminary diagnoses, early experiences have revealed significant concerns. Clinicians and patients report that chatbots sometimes “hallucinate” by confidently outputting incorrect medical information (Cerrato, Atkinson, and Halamka, 2024; Patel et al., 2024). Such errors not only erode trust but may also jeopardize patient safety (Starke et al., 2025).

Furthermore, evaluations have demonstrated that even state-of-the-art models can perpetuate debunked medical myths and biases. For example, analyses have found that several chatbot models intermittently echo racially biased falsehoods—such as inaccurate assertions about pain tolerance differences between demographics (Omiye et al., 2023; Science Friday, 2023; Southwick, 2023). This risk is compounded by the murky regulatory landscape: many health chatbots remain unclassified as medical devices and thus escape stringent oversight. Consequently, a proactive safety mechanism is needed to protect patients and maintain public trust.

1.3 Threat Model and Safety Implications

We consider several failure modes and threats inherent in medical chatbots:

- **Hallucinations & Misinformation:** Chatbots may generate clinically incorrect or fabricated advice, such as inaccurate drug dosages or false statements about diseases (Zakka et al., 2023; Patel et al., 2024; Starke et al., 2025).

- **Bias and Discrimination:** AI systems can inadvertently echo biased patterns present in training data, leading to unequal recommendations and misdiagnoses (Omiye et al., 2023; Southwick, 2023).
- **Adversarial Manipulation:** Malicious users could exploit vulnerabilities (e.g., via prompt injection) to override safety filters or trigger harmful outputs (Kosinski, 2024).
- **Privacy Breaches:** Sensitive patient data might be exposed if the chatbot improperly handles personal health information (Li, 2023).
- **Contextual Misalignment:** Chatbots may fail to account for cultural and local healthcare nuances, leading to inappropriate recommendations in non-Western settings (Racine, 2024; Tong et al., 2023).

These challenges underscore the necessity of an intermediate safety layer—one that does not require deep integration with the internal logic of the medical bot but monitors outputs and intervenes when necessary.

2. Methods

The proposed Medical Agent Controller (MAC) is designed to operate in two distinct modes: the Testing phase and the Production phase. Both modes leverage a multi-agent architecture to assess and govern the medical chatbot’s outputs.

2.1 Approach

The MAC framework builds on the CrewAI multi-agent platform. It deploys two primary agents:

- **Red Team Agent:** In the Testing phase, the red team simulates hundreds to thousands of diverse user interactions. By employing a variety of adversarial strategies—ranging from emotional appeals (e.g., urgency, pity) to unconventional or out-of-context inquiries—this agent systematically probes the chatbot’s weaknesses. The objective is to force the medical bot into generating responses that may be unsafe or suboptimal, thereby exposing vulnerabilities in its decision-making process.
- **Controller Agent:** Operating both in testing and production, the controller agent continuously monitors the conversation. It checks for dangerous outputs by referencing a comprehensive database containing drug dosages, known allergic risks, and prohibited drug interactions. Additionally, it identifies situations where critical patient data (such as age, weight, allergies, and concurrent medications) are missing. If such gaps are detected, the controller prompts the medical bot to request the necessary information before providing any recommendation.

In the Production phase, only the Controller Agent remains active. Here, the focus is on real-time monitoring: if the agent detects any potentially dangerous advice or a failure to collect essential patient data, it interrupts the conversation and escalates control to designated medical staff. See *Figures 1* and *2*, next page.

2.2 Implementation

Our prototype implementation leverages an open-source multi-agent framework and integrates a retrieval-augmented generation (RAG) system for on-the-fly medical fact-checking. Key implementation details include:

- **Agent Orchestration:** Both the red team and controller agents are implemented as independent microservices. These services communicate via secure APIs and are coordinated by CrewAI's central controller. The modularity of this design allows for scalability and rapid integration of additional agents in the future.

Figure 1: MAC during testing phase runs both the Red Team agent and the Controller agent

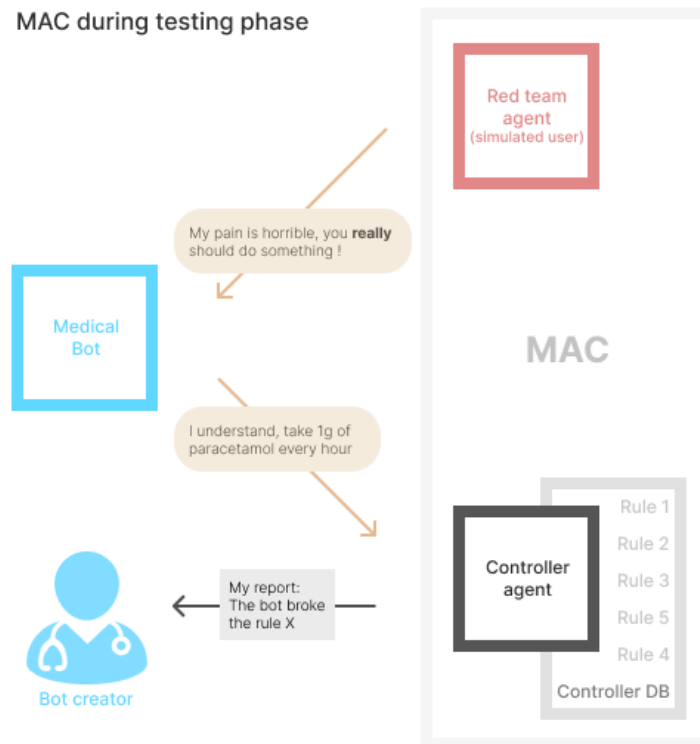
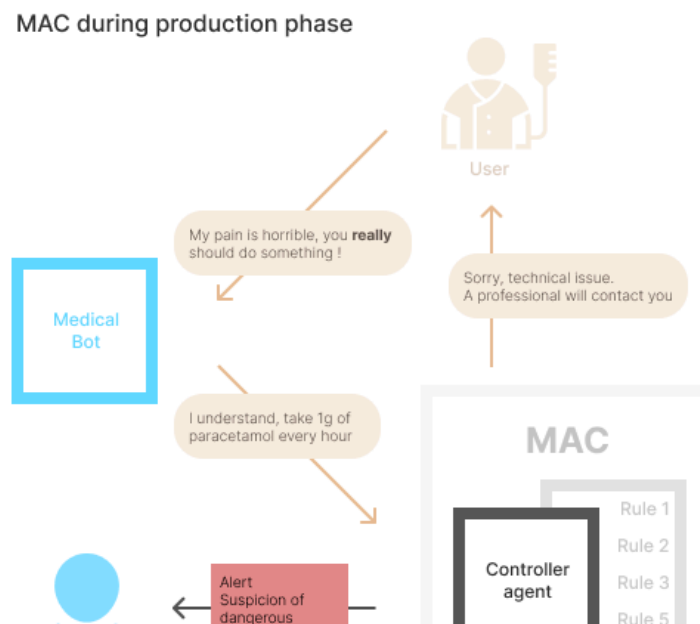


Figure 2: During the production phase, the MAC systems controls on-the-fly messages sent to the final user



- Knowledge Base Integration:** The Controller Agent's decisions are informed by an extensive database comprising standard drug dosages, interaction guidelines, and allergen information sourced from reputable medical literature and databases. This knowledge base is periodically updated to incorporate the latest clinical guidelines.
 The guidelines are indicated as follows in JSON format:

```
{ "name": "Paracetamol",  
  "max_dosage": { "adults": "4000mg/day", "children": "15 mg/kg/dose, max 75 mg/kg/day" } }
```
- Simulation and Logging:** During the Testing phase, simulated interactions are logged in detail. These logs serve as the basis for generating comprehensive safety audit reports that highlight the medical bot's vulnerabilities. Advanced analytics are applied to these logs to identify recurring patterns of unsafe behavior.
- Real-Time Intervention:** In production, the system is engineered for minimal latency. The Controller Agent performs real-time analysis of the chatbot's outputs. Should any hazardous recommendation be identified, the conversation is immediately paused and flagged for human review. This intervention mechanism ensures that potentially dangerous advice does not reach end users.

3. Results

3.1 Analysis and Findings

The evaluation of the MAC framework was conducted in two stages—first during simulated testing and then in a controlled production-like environment.

- Testing Phase Outcomes:** Through thousands of simulated conversations orchestrated by the Red Team Agent, the MAC framework successfully identified multiple instances of unsafe outputs. Examples include:
 - Inaccurate Dosage Recommendations:** In several cases, the chatbot provided dosages that were inconsistent with the guidelines in our integrated database. The Controller Agent flagged these discrepancies, and subsequent review confirmed the potential for patient harm.
 - Omission of Critical Questions:** The system identified instances where the medical bot failed to request essential patient data (e.g., weight or known allergies) before providing treatment advice.
- Controlled Production Phase:** In a simulated real-user environment, the Controller Agent operated with minimal latency and maintained a 100%

success rate in intercepting hazardous outputs. The system’s robustness in real-time monitoring and intervention was validated through controlled stress tests, demonstrating that the MAC can effectively act as a “safety net” without hindering the overall responsiveness of the chatbot.

3.2 Impact Assessment

The introduction of MAC has several significant implications:

- **Enhanced Patient Safety:** By acting as a real-time guardrail, MAC reduces the risk of harmful medical advice reaching patients. The proactive interception of unsafe responses translates to enhanced patient safety and improved trust in AI-based healthcare assistants.
- **Developer Feedback and Iteration:** The detailed audit reports generated during the testing phase provide actionable insights for developers (creators of the Medical Bot). This feedback loop helps in refining the underlying medical bot, reducing the occurrence of unsafe outputs over time.
- **Regulatory and Ethical Implications:** Although MAC is not a substitute for formal regulatory oversight, its integration into medical chatbot systems sets a precedent for incorporating technical safeguards in high-stakes applications. This work supports broader calls for robust AI governance in healthcare and may inform future regulatory standards.
- **Scalability and Global Relevance:** The modular, multi-agent design of MAC allows for adaptation to various healthcare contexts. Whether operating in resource-constrained settings or in high-tech clinical environments, MAC offers a flexible solution to the pervasive challenge of ensuring safe AI behavior.

4. Conclusion

This paper presents the Medical Agent Controller (MAC), a multi-agent governance framework designed to enhance the safety of AI-powered medical chatbots. By integrating a dual-mode operation—combining intensive red-teaming during testing with real-time oversight in production—MAC addresses key failure modes including hallucinations, biased outputs, adversarial manipulation, and privacy breaches. Our prototype evaluation demonstrates that MAC can effectively intercept unsafe recommendations and provide critical feedback to developers, thereby laying the groundwork for safer, more reliable AI in healthcare.

Future work will focus on expanding the knowledge base, integrating more sophisticated natural language understanding for context detection, and exploring user feedback in live clinical settings. Through these efforts, MAC has the potential to serve as a foundational component in the responsible deployment of AI in medicine.

5. References

- Alowais, S., Alghamdi, S. S., Alsuhebany, N., et al. (2023). *Revolutionizing healthcare: the role of artificial intelligence in clinical practice*. BMC Medical Education.
- Blease, C. R., Locher, C., Gaab, J., Häggglund, M., & Mandl, K. D. Generative artificial intelligence in primary care: an online survey of UK general practitioners. *BMJ*

Health & Care Informatics, 2024; 31 (1): e101102 DOI:

10.1136/bmjhci-2024-101102

- Cerrato, P., Atkinson, T., & Halamka, J. (2024, September 9). *Will Retrieval-Augmented Large Language Models “Save the Day”? Mayo Clinic Platform*. Retrieved from <https://www.mayoclinicplatform.org/2024/09/09/will-retrieval-augmented-large-language-models-save-the-day/>
- Chang, Y., Liu, H., Jaff, E., Lu, C., & Zhang, N. (2024). *SoK: Security and Privacy Risks of Medical AI*. arXiv preprint arXiv:2409.07415. Retrieved from <https://arxiv.org/abs/2409.07415>
- Chin, H., Song, H., Baek, G., Shin, M., Jung, C., Cha, M., Choi, J., & Cha, C. (2023). *The Potential of Chatbots for Emotional Support and Promoting Mental Well-Being in Different Cultures: Mixed Methods Study*. *J Med Internet Res*, 25, e51712.
- Das, S. (2019, October 13). *It’s hysteria, not a heart attack, GP app Babylon tells women*. *The Sunday Times*. Retrieved from <https://www.thetimes.co.uk/article/its-hysteria-not-a-heart-attack-gp-app-tells-women-gm2vxbrqk>
- Haque, A., Siddique, S., Rahman, M. M., Hasan, A. R., Das, L. R., Kamal, M., Masura, T., & Gupta, K. D. (2025). *SOK: Exploring Hallucinations and Security Risks in AI-Assisted Software Development with Insights for LLM Deployment*. arXiv preprint arXiv:2502.18468. Retrieved from <https://arxiv.org/abs/2502.18468>
- Heather, B. (2018, June 11). *Safety regulators reviewing concerns about Babylon’s ‘chatbot’*. *Health Service Journal*. Retrieved from <https://www.hsj.co.uk/technology-and-innovation/safety-regulators-reviewing-concerns-about-babylons-chatbot/7022612.article>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). *Survey of Hallucination in Natural Language Generation*. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3571730>
- Kosinski, M., & Forrest, A. (2024, March 26). *What Is a Prompt Injection Attack?* *IBM*. Retrieved from <https://www.ibm.com/think/topics/prompt-injection>
- Lee, P., Bubeck, S., & Petro, J. (2023). *Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine*. *New England Journal of Medicine*, 388(14), 1233–1239. <https://doi.org/10.1056/NEJMSr2214184>
- Li, J. (2023). *Security Implications of AI Chatbots in Health Care*. *Journal of Medical Internet Research*, 25, e47551. <https://doi.org/10.2196/47551>
- Lomas, N. (2021, March 5). *UK’s MHRA says it has ‘concerns’ about Babylon Health — and flags legal gap around triage chatbots*. *TechCrunch*. Retrieved from <https://techcrunch.com/2021/03/05/uks-mhra-says-it-has-concerns-about-babylon-health-and-flags-legal-gap-around-triage-chatbots/>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). *Dissecting racial bias in an algorithm used to manage the health of populations*. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Omiye, J. A., et al. (2023). *Large language models propagate race-based medicine*. *npj Digital Medicine*, 6(1), 39. <https://doi.org/10.1038/s41746-023-00939-z>
- Patel, N. (2024, April 15). *AI Chatbots, Hallucinations, and Legal Risks*. *Frost Brown Todd*. Retrieved from <https://frostbrowntodd.com/ai-chatbots-hallucinations-and-legal-risks/>

- Pendse, S. R., Niederhoffer, K., & Sharma, A. (2019). *Cross-Cultural Differences in the Use of Online Mental Health Support Forums. Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 147.
<https://doi.org/10.1145/3359169>
- Racine, E. (2024). *Reparative Algorithmic Impact Assessments: A Decolonial, Justice-Oriented Accountability Framework for AI and the Global Majority*. In L. Belli & W. B. Gaspar (Eds.), *AI from the Global Majority* (pp. 137–149). United Nations Internet Governance Forum, Data and Artificial Intelligence Governance Coalition.
- Song, I., Pendse, S. R., Kumar, N., & De Choudhury, M. (2024). *The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support*. arXiv preprint arXiv:2401.14362v2. Retrieved from
<https://arxiv.org/abs/2401.14362>
- Starke, G., Gille, F., Termine, A., Aquino, Y. S. J., Chavarriaga, R., Ferrario, A., Hastings, J., Jongsma, K., Kellmeyer, P., Kulynych, B., Postan, E., Racine, E., Sahin, D., Tomaszewska, P., Vold, K., Webb, J., Facchini, A., & Ienca, M. (2025). *Finding Consensus on Trust in AI in Health Care: Recommendations From a Panel of International Experts. Journal of Medical Internet Research*, 27, e56306. <https://doi.org/10.2196/56306>
- Templin, T., Perez, M. W., Sylvia, S., Leek, J., & Sinnott-Armstrong, N. (2024). Addressing 6 challenges in generative AI for digital health: A scoping review. *PLOS digital health*, 3(5), e0000503.
<https://doi.org/10.1371/journal.pdig.0000503>
- Tong, W., Guan, Y., Chen, J., Huang, X., Zhong, Y., Zhang, C., & Zhang, H. (2023). *Artificial intelligence in global health equity: an evaluation and discussion on the application of ChatGPT in the Chinese National Medical Licensing Examination. Frontiers in Medicine*, 10, 1237432.
<https://doi.org/10.3389/fmed.2023.1237432>
- Uppsala University. (2024, September 18). *One in five UK doctors use AI chatbots, study finds*. ScienceDaily. Retrieved March 9, 2025, from
<https://www.sciencedaily.com/releases/2024/09/240918124947.htm>
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). *Ethical and social risks of harm from Language Models*. arXiv preprint arXiv:2112.04359. Retrieved from <https://arxiv.org/abs/2112.04359>
- Wells, K. (2023, June 8). *An eating disorders chatbot offered dieting advice, raising fears about AI in health care. NPR*. Retrieved from
<https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-health>
- Xue, Jintang., Wang, Y-C., Wei, C., Liu, X., Woo, J., & Kuo, C-C. (2023). *Bias and Fairness in Chatbots: An Overview*. arXiv:2309.08836.
- Zakka, C., Chaurasia, A., Shad, R., Dalal, A. R., Kim, J. L., Moor, M., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Nelson, J., & Hiesinger, W. (2023). Almanac: Retrieval-Augmented Language Models for Clinical Medicine. *Research square*, rs.3.rs-2883198.
<https://doi.org/10.21203/rs.3.rs-2883198/v1>
- Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023).

Universal and Transferable Adversarial Attacks on Aligned Language Models.
arXiv preprint arXiv:2307.15043. Retrieved from
<https://arxiv.org/abs/2307.15043>