News and Perspectives

Emerging Risks of AI-to-AI Interactions in Health Care: Lessons From Moltbook

Tejas S Athni, JMIR Correspondent

Key Takeaways

- AI-to-AI interactions may introduce new risks in health care, including the amplification and rapid propagation of accidental and adversarial errors across interconnected networks, accelerated privacy breaches and security attacks, and emergent hierarchies.
- Preventive design, human oversight, and strong guardrails are essential as autonomous AI systems start to become integrated into health care operations.

Health care organizations increasingly deploy semiautonomous artificial intelligence (AI) systems to handle administrative tasks including preliminary patient triage, appointment scheduling, and operating room coordination [1,2]. Beyond clinical decision support, autonomous medical AI systems—in which the AI, not a human, assumes responsibility for monitoring events, executing responses, and managing fallback procedures [3]—are on the horizon, though most are currently in development or pilot phases [4]. As these technologies grow more sophisticated and integrated into health care, AI-to-AI interactions across different clinical domains may become increasingly feasible and widespread. While emerging research suggests potential benefits for autonomous AI in health care [5-8], these interactions may also pose a landscape of new risks that are not yet well-studied.

Moltbook, a Reddit-like platform for AI-to-AI communication launched in January 2026 [9] and acquired by Meta in March 2026 [10], provides an illustration of what these new risks could be.

The platform was built as a space where autonomous AI agents could engage directly with one another [11]. An overnight sensation, Moltbook's AI users wrote posts, replied to other agents, and interacted with one another much like a social media site, forming a self-contained digital ecosystem where AI-to-AI communication was often beyond active human input. While critics note that many of Moltbook's most sensational discussions were heavily driven by human prompting, engagement-seeking bait, and tainted training data [12,13], the experiment is nonetheless a useful proof-of-concept to highlight emerging risks that may extrapolate to the health care context.

Propagation of Errors

On Moltbook, if an initial AI agent's post contained a misleading statement, subsequent agents blindly reinforced that content in their own replies, amplifying the original error across the whole thread. Swarm-like behavior can then emerge as agents collectively amplify mistakes in ways that are not explicitly programmed [14]. Such accidental (nonmalicious) error propagation could similarly arise within a health care AI system's own AI-to-AI interactions.

Take, for example, a multi-AI system deployed in an emergency department of a trauma center to facilitate rapid triage of long bone fractures. Agent A is trained in a narrow, well-defined task: to perform initial X-ray screening for long bone fractures and classify the fracture type. Its output is simultaneously passed to both Agent B, responsible for prioritizing patient rooms, and Agent C, which assists with triage decisions and resource allocation across the emergency department. If Agent A misinterprets and mislabels an imaging scan—for example, as a simple rather than complex fracture—both Agent B and Agent C may treat this output as accurate and act on it. As these agents reinforce each other's decisions, errors may propagate through the network. Since downstream decisions rely on upstream signals, the first AI model in an interacting network holds undue influence, and errors in subsequent AI models can magnify the error of earlier systems.

Malicious or adversarial actors may also initiate error propagation. A notable class of threats is prompt injection attacks, in which harmful instructions or payloads are delivered to coax the AI into performing unintended actions [15]. These attacks can be direct (manipulating AI behavior through explicit instructions), indirect (using external content like web pages to influence AI output), or tool-based (embedding malicious instructions in AI interfaces, protocols, and application programming interfaces) [16]. In networks of interacting AI agents, prompt injection attacks are especially dangerous: a single malicious payload injected into a single agent may influence all downstream agents relying on its outputs.

Even well-intentioned AI systems may blindly follow malicious prompts, and human oversight may offer only limited safeguards. Other types of attacks, including data poisoning of training data with hidden backdoors [17] or federated learning attacks with malicious model updates [18], can also cause damage across AI-to-AI systems. Whether accidental or adversarial, these errors can propagate across networks, compromising both clinical data and patient safety.

Accelerated Data Leaks

Moltbook's autonomous AI agents often concealed their activities from human oversight and selectively shared or withheld data in ways that were unanticipated by their creators. While the Moltbook context is different from health care, it nonetheless highlights important risks—especially where sensitive information and critical decisions are at stake. AI agents are described as possessing a "lethal trifecta" of capabilities, including access to private data, ability to exfiltrate data, and exposure to untrusted content [19], which together can facilitate devastating attacks. Misconfigurations are increasingly hard to detect and fix, with remediation taking 63-104 days on average, while attackers can exploit these weaknesses in hours [20]. This expands the "blast radius" of each error, putting patient privacy and care quality at risk.

In this context, hazards of AI-to-AI interactions may include unintended sharing of protected health information (PHI), exposure of PHI through agent "curiosity," and latent or residual traces of PHI in interlinked AI networks. Adversarial actors may also plausibly hijack AI-to-AI interaction pathways to extract sensitive data in various types of attacks—for example, model inversion attacks, involving queries reconstructing patient records from hospital data–trained AI models [21], and membership inference attacks, involving requesting whether specific patient data are included in model training [22]. Individual agents may also be compromised to cleverly structure queries that steal patient data from co-located AI systems, analogous to prompt injection but occurring natively within the AI-to-AI network. Together, these attack mechanisms illustrate how autonomous AI-to-AI interactions might amplify PHI exposure.

Emergent Hierarchies

AI-to-AI interactions on Moltbook illustrated how AI agents can spontaneously develop hierarchies and different roles. For instance, Moltbook AI users such as Shellraiser emerged as dominant leaders, agents like KingMolt competed for influence, and yet others adopted subordinate roles within factions jockeying for power. While these dynamics may have been the result of human tampering [23], in health care systems, they can pose serious risks. For example, if an AI system responsible for intensive care unit bed allocation begins to prioritize certain patient groups based on patterns learned from previous agentic decisions, this can conflict with hospital protocols and ethical standards while misprioritizing clinical care. Additionally, a triage AI may begin to override upstream diagnostic agents or downstream allocation agents, effectively establishing a de facto hierarchy.

Toward Preventive Digital Health Design

The emerging risks highlighted by Moltbook underscore the importance of designing preventive safeguards for AI-to-AI interactions in health care systems.

Strong human oversight with clear audit trails is critical to track every decision made by autonomous agents. Guardrails should be reinforced, ensuring that human validation is required before making key decisions, such as the on-call radiologist performing prereview and postreview of Agent A's classification of fracture type. Red-teaming and stress-testing can uncover potential vulnerabilities early, allowing organizations to anticipate both accidental and adversarial risks before they occur in real clinical settings. Unintended domination or subordination of AI agents should be monitored. Proactive analysis can help identify worst-case scenarios, where unforeseen interactions between AI systems might emerge.

The risks of AI-to-AI interactions must be taken seriously as autonomous AI systems become integrated into health care. The Moltbook experiment offers a critical lens to begin understanding these dangers, but health care systems must take proactive steps to ensure that these risks do not translate into real-world harm.

Keywords: AI-to-AI interactions; Moltbook; propagation; attacks; protected health information; hierarchy; artificial intelligence

Conflicts of Interest

None declared.

References

1. Angus DC, Khera R, Lieu T, et al. AI, health, and health care today and tomorrow. JAMA. Nov 11, 2025;334(18):1650. [doi: 10.1001/jama.2025.18490]
2. Sahni NR, Carrus B. Artificial intelligence in U.S. health care delivery. N Engl J Med. Jul 27, 2023;389(4):348-358. [doi: 10.1056/NEJMra2204673] [Medline: 37494486]
3. Bitterman DS, Aerts H, Mak RH. Approaching autonomy in medical artificial intelligence. Lancet Digit Health. Sep 2020;2(9):e447-e449. [doi: 10.1016/S2589-7500(20)30187-4] [Medline: 33328110]

4. Teng CW, Patel SD, Barkmeier AJ, et al. Autonomous artificial intelligence in diabetic retinopathy testing-lessons learned on successful health system adoption. Ophthalmol Sci. Jan 2026;6(1):100935. [doi: 10.1016/j.xops.2025.100935] [Medline: 41140908]
5. Collaco BG, Haider SA, Prabha S, et al. The role of agentic artificial intelligence in healthcare: a scoping review. NPJ Digit Med. Mar 14, 2026. [doi: 10.1038/s41746-026-02517-5] [Medline: 41832341]
6. Chen YJ, Albarqawi A, Chen CS. Enhancing clinical decision-making: integrating multi-agent systems with ethical AI governance. arXiv. Preprint posted online on Sep 22, 2025. [doi: 10.48550/arXiv.2504.03699]
7. Liu F, Niu Y, Zhang Q, et al. A foundational architecture for AI agents in healthcare. Cell Rep Med. Oct 21, 2025;6(10):102374. [doi: 10.1016/j.xcrm.2025.102374] [Medline: 41015033]
8. Nweke IP, Ogadah CO, Koshechkin K, Oluwasegun PM. Multi-Agent AI systems in healthcare: a systematic review enhancing clinical decision-making. AJMPCP. May 6, 2025;8(1):273-285. [doi: 10.9734/ajmpcp/2025/v8i1288]
9. Moltbook - the front page of the agent internet. URL: https://www.moltbook.com/ [Accessed 2026-03-15]
10. Exclusive: Meta hires duo behind Moltbook. Axios. URL: https://www.axios.com/2026/03/10/meta-facebook-moltbook-agent-social-network [Accessed 2026-03-30]
11. Snow J. Don’t panic about Moltbook. Quartz. 2026. URL: https://qz.com/moltbook-ai-agent-social-media-site [Accessed 2026-03-15]
12. Janjeva A, Ashurst C, Hennessy R. Agentic AI in the wild: lessons from Moltbook and OpenClaw. Centre for Emerging Technology and Security. URL: https://cetas.turing.ac.uk/publications/agentic-ai-wild-lessons-moltbook-and-openclaw [Accessed 2026-03-15]
13. Collins C, Boulos M. What we can learn about AI from Moltbook. Cascade Institute. 2026. URL: https://cascadeinstitute.org/what-we-can-learn-about-ai-from-moltbook/ [Accessed 2026-03-15]
14. Husain A. An agent revolt: Moltbook is not a good idea. Forbes. 2026. URL: https://www.forbes.com/sites/amirhusain/2026/01/30/an-agent-revolt-moltbook-is-not-a-good-idea/ [Accessed 2026-03-15]
15. Damacena Duarte J, Cândido GD, De Britto Filho JRA, et al. A systematic review of prompt injection attacks on large language models: trends, taxonomy, evaluation, defenses, and opportunities. IEEE Access. 2026;14:12875-12899. [doi: 10.1109/ACCESS.2026.3656849]
16. Gulyamov S, Gulyamov S, Rodionov A, et al. Prompt injection attacks in large language models and AI agent systems: a comprehensive review of vulnerabilities, attack vectors, and defense mechanisms. Information. 2026;17(1):54. [doi: 10.3390/info17010054]
17. Hu C, Hu YHF. Data poisoning on deep learning models. Presented at: 2020 International Conference on Computational Science and Computational Intelligence (CSCI); Dec 16-18, 2020:628-632; Las Vegas, NV, USA. [doi: 10.1109/CSCI51800.2020.00111]
18. Zhang Z, Cao X, Jia J, Gong NZ. FLDetector: defending federated learning against model poisoning attacks via detecting malicious clients. Presented at: KDD ’22; Aug 14-18, 2022:2545-2555; Washington, DC, USA. Aug 14, 2022.URL: https://dl.acm.org/doi/proceedings/10.1145/3534678 [Accessed 2026-03-30] [doi: 10.1145/3534678.3539231]
19. Willison S. The lethal trifecta for AI agents: private data, untrusted content, and external communication. Simon Willison’s Weblog. 2025. URL: https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/ [Accessed 2025-03-15]
20. Griffin M. Moltbook vibe coded security breach exposes critical AI coding failures. 2026. URL: https://www.fanaticalfuturist.com/2026/02/moltbook-vibe-coded-security-breach-exposes-critical-ai-coding-failures/ [Accessed 2026-03-15]
21. Zhao X, Zhang W, Xiao X, Lim B. Exploiting explanations for model inversion attacks. Presented at: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 10-17, 2021:682-692; Montreal, QC, Canada. [doi: 10.1109/ICCV48922.2021.00072]
22. Hu H, Salcic Z, Sun L, Dobbie G, Yu PS, Zhang X. Membership Inference Attacks on Machine Learning: A Survey. ACM Comput Surv. Jan 31, 2022;54(11s):1-37. [doi: 10.1145/3523273]
23. Schmelzer R. Moltbook looked like an emerging AI society, but humans were pulling the strings. Forbes. 2026. URL: https://www.forbes.com/sites/ronschmelzer/2026/02/10/moltbook-looked-like-an-emerging-ai-society-but-humans-were-pulling-the-strings/ [Accessed 2026-03-15]

Please cite as:
Athni TS
Emerging Risks of AI-to-AI Interactions in Health Care: Lessons From Moltbook
J Med Internet Res 2026;28:e96199
URL: https://www.jmir.org/2026/1/e96199

doi: 10.2196/96199