Infection and Drug Resistance

Dovepress
Taylor & Francis Group

Open Access Full Text Article

PERSPECTIVES

A Theoretical Framework for *Shared Reasoning Fragility* in Clinician-Chatbot Interactions Through the Example of Antibiotic Prescribing

Daniele Roberto Giacobbe[1,2], Alessandra Agnese Grossi[3,4], Cristina Marelli[5,6], Marco Muccio[1], Sabrina Guastavino[7], Ylenia Murgia[8], Sara Mora[9], Alessio Signori[10,11], Nicola Rosso[9], Mauro Giacomini[8], Cristina Campi[7,12], Michele Piana[7,12], Matteo Bassetti[1,2]

[1]UO Clinica Malattie Infettive, IRCCS Azienda Ospedaliera Metropolitana, Genoa, Italy; [2]Department of Health Sciences (DISSAL), University of Genoa, Genoa, Italy; [3]Department of Human Sciences and Innovation for the Territory, University of Insubria, Varese, Italy; [4]Department of Biotechnologies and Life Sciences, Center for Clinical Ethics, University of Insubria, Varese, Italy; [5]CESP - INSERM U1018, Oncostat, Labeled Ligue Contre le Cancer, Gustave Roussy, Université Paris-Saclay, Villejuif, France; [6]Institut Curie - INSERM U1331, Team Statistics Applied to Personalized Medicine, Paris, France; [7]Department of Mathematics (DIMA), University of Genoa, Genoa, Italy; [8]Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS), University of Genoa, Genoa, Italy; [9]UO Information and Communication Technologies, IRCCS Azienda Ospedaliera Metropolitana, Genoa, Italy; [10]Department of Health Sciences (DISSAL), Section of Biostatistics, University of Genoa, Genoa, Italy; [11]Ospedale Policlinico San Martino, IRCCS Azienda Ospedaliera Metropolitana, Genoa, Italy; [12]Life Science Computational Laboratory (LISCOMP), IRCCS Azienda Ospedaliera Metropolitana, Genoa, Italy

Correspondence: Daniele Roberto Giacobbe, Department of Health Sciences (DISSAL), University of Genoa, Via A. Pastore 1, Genoa, 16132, Italy, Email danieleroberto.giacobbe@unige.it

Abstract: General-purpose large language model (LLM)-based chatbots are increasingly used by clinicians to discuss medical problems, including antibiotic prescribing. Their use creates an unprecedented setting for clinical reasoning in which diagnostic and therapeutic thinking becomes dynamically shared between human and machine. Here, we propose a theoretical framework, intended for subsequent empirical assessment, around the concept of *shared reasoning fragility*, defined as the potential instability arising from the interaction between clinician reasoning and the chatbot's opaque, association-based processes, which are structurally different from classical human reasoning. The theoretical framework is based on the conceptual argument that, while the black box dilemma is often discussed for classification-oriented clinical decision support systems with an emphasis on explainability versus external validation, chatbot-assisted practice introduces a distinct problem: chatbots can accompany clinicians throughout the entire reasoning pathway rather than being consulted only at the final decision point. In the present perspective, we argue more explicitly that the fragility of this continuous co-reasoning primarily stems from its novelty and pervasiveness. Using strictly illustrative examples in antibiotic prescribing, we suggest the theoretical possibility that fluent and convincing outputs may redirect attention, mask omissions in work-up, and subtly shift hypothesis selection during shared clinical reasoning processes. While it is important to stress that our framework is purely theoretical and thus cannot be confirmed at the present stage, our considerations are intended to motivate the required quantitative research to confirm or refute *shared reasoning fragility*, measure its extent, and evaluate downstream implications for patient care.

Keywords: healthcare, infection, antibiotic prescribing, artificial intelligence, machine learning, deep learning, natural language processing

Introduction

ChatGPT 3.5 (OpenAI), a general-purpose large language model (LLM)-based tool that allows users to chat with a non-human counterpart in an unprecedentedly fluent human language, was released to the public on 30 November 2022.[1] This release marked the beginning of broader public discovery and use of both free and subscription versions of ChatGPT 3.5 and subsequent versions (up to 5.2 at the time of writing), as well as other LLM-based chatbots such as Gemini (Google) and Claude (Anthropic), among others.

Following this development, for example, the number of weekly active ChatGPT users on consumer plans (Free, Plus, Pro) has grown exponentially over the past two years, approaching 800 million in September 2025.[2] The Gemini app has similarly

https://doi.org/10.2147/IDR.S601126
Received: 4 February 2026
Accepted: 4 May 2026
Published: 11 May 2026

been reported to exceed 650 million monthly users, while Claude was reported in 2025 to have 18.9 million monthly active website users and 2.9 million monthly active app users, with substantial growth between early and late 2024.[3,4] Given these figures and the ease of access such tools via smartphones and computers, clinicians are likely among the millions of users discussing medical matters with frontier chatbots daily. This does not necessarily imply inappropriate chatbot use or breaches of patient privacy, as simulated cases and general medical and research questions can be discussed without disclosing identifiable patient data or violating current laws or regulations. Consistently, a growing number of research has examined the potential uses and performance of chatbots, both independently and as decision-support tools for clinicians, in diagnostic and therapeutic tasks, including antibiotic prescribing.[5–15]

Using chatbots for these purposes creates a novel context for clinical reasoning, in which thinking - from the consideration of symptoms and signs through diagnosis and treatment - is no longer exclusively human but is dynamically shared with a machine. This interaction may introduce instability between chatbot outputs, generated through opaque associative patterns structurally distinct from classical human reasoning, and clinicians' own reasoning processes, giving rise to an insufficiently characterized mode of decision-making. We have previously introduced and briefly defined this interaction as *shared reasoning fragility*.[16] The present perspective develops this concept into a more explicit theoretical framework for healthcare, using antibiotic prescribing as an illustrative case. Notably, it is possible that the framework may also apply to domain-specific LLMs; however, we focus here on general-purpose LLMs, as they likely represent the most widely used tools at present.

The Black Box Dilemma: The Nuanced Different Angle of Human-Chatbot Interactions vs Classification Tasks

One of the most debated issues in applying artificial intelligence (AI) to healthcare is the black box nature of some predictive algorithms. This refers to the partial or complete inscrutability of the internal calculations and learned associations through which some models use data to predict clinical events, such as diagnosing a particular infection or selecting empirical antibiotic therapy in septic shock.[17,18] This concern also applies to general-purpose LLMs, which underlie contemporary chatbot applications, because they similarly rely on "black box" architectures to predict the next token, meaning a word or part of a word, in response to a user query.[7,19–22] For example, as shown in Figure 1, such a query may concern whether an antibiotic should be administered in a given clinical scenario and, if so, which one(s).

For non-LLMs black box predictive algorithms used in classification tasks, such as predicting the presence of infection due to a multidrug-resistant organism, the debate has largely focused on the balance between explainability and external validation. Explainability relates to the field of explainable AI (XAI), in which a black box model's prediction is "explained" by a more interpretable surrogate model.[23–27] Conceptually, such models are typically less accurate than the underlying black box; otherwise, they would replace it. This creates a well-recognized tension: explanations may help identify omissions, misleading associations, or biases, but, as approximations of a more complex process, they may also be unreliable. For this reason, some experts prioritize external validation over explainability.[28] However, this position entails accepting at least partial non-explainability: some reasons for a given prediction remain inaccessible in exchange for improved performance. Whether, and to what extent, such a compromise is acceptable remains debated and likely depends on the specific clinical task. Importantly, accepting reduced explainability also raises ethical concerns, particularly regarding informed consent (Table 1).[29–35]

While some of these considerations also apply to next-token prediction by general-purpose LLMs, we argue that chatbot outputs raise distinct conceptual issues that are central to the present framework. First, in our view, the explanatory techniques described above are not conceptually applicable to chatbot responses. Providing references and sources can help ground claims by identifying their evidentiary basis, but it does not technically explain why the chatbot generated a particular response from that evidence. Similarly, prompting the chatbot to generate chain-of-thought reasoning may improve performance and reduce hallucinations,[36,37] but it does not necessarily explain why a given output was produced. Chain-of-thought reasoning may offer an initial mechanistic window into how LLMs operate by modulating task-dependent node activation,[38] but what these activated nodes do - including which hidden features they construct or how semantic and meta-semantic information is integrated - remains largely unknown. Furthermore, the

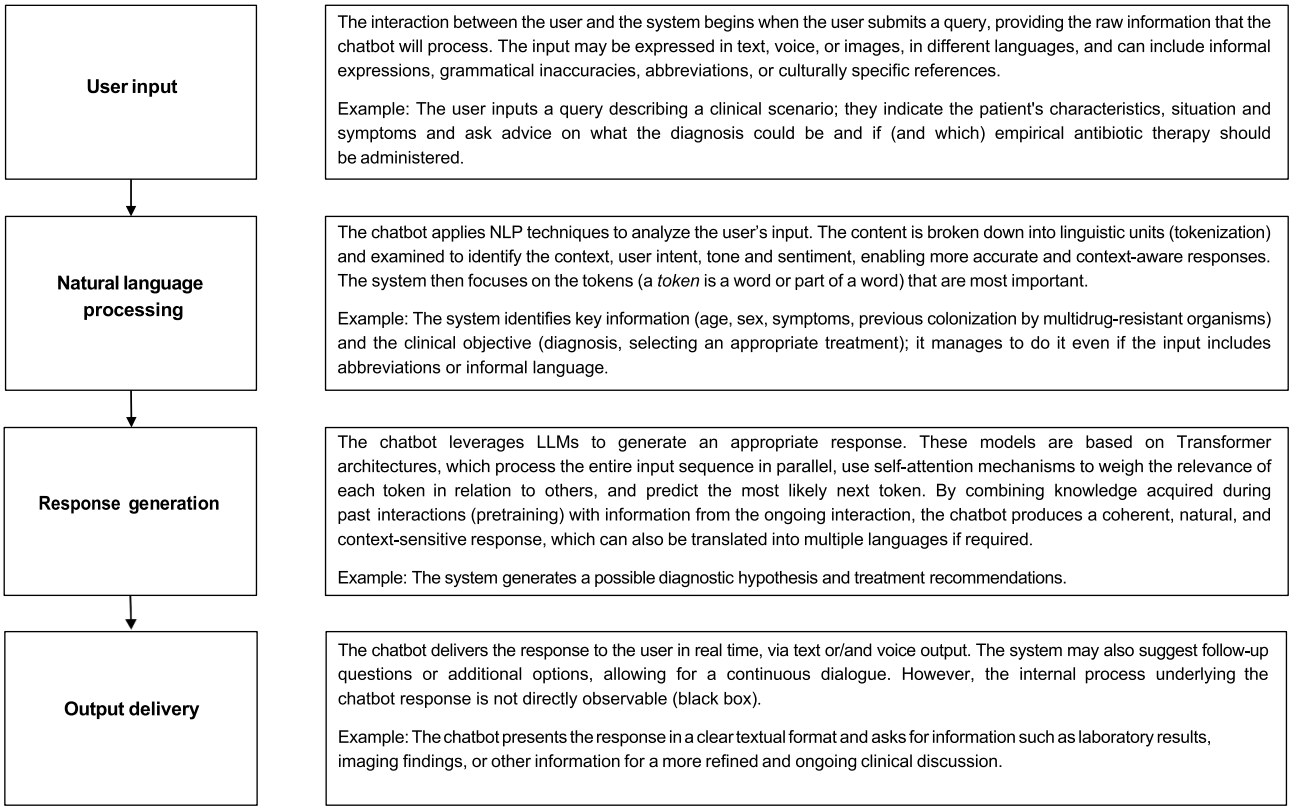

Figure 1 Workflow of the interaction between clinicians and LLMs-based chatbots.
Abbreviations: NLP, natural language processing; LLMs, large language models.

reliability of chain-of-thought reasoning must be interpreted in light of the broader instability of LLM outputs. Studies on persistent instability in LLMs personality measurements suggest that model responses are highly sensitive to prompt phrasing, reasoning style, and conversational context.[39] This raises concerns about whether chain-of-thought outputs can be interpreted as stable reflections of underlying reasoning. In parallel, recent work on chain-of-thought monitorability recognizes that reasoning traces may make model behaviour partly observable, but their interpretation remains context-dependent and may change as models are updated or optimized differently.[40]

Table 1 Key Ethical Issues Related to Informed Consent in AI-Supported Decision-Making

Ethical Issue	Relevance to Informed Consent
Opacity and limited interpretability of AI systems	Undermined patient's ability to understand clinical decisions and treatment options when AI systems lack transparency or generate outputs that clinicians themselves struggle to interpret, thereby compromising both explanation and IC.
Bias and lack of representativeness in training and validation data	Undermined patient's ability to make informed choices when algorithmic limitations and potential discriminatory effects are not transparent, exposing individuals to unequal risks without their awareness.
Secondary use of patient data for algorithm development	Undermined patient's ability to exercise autonomy over personal data when secondary uses for algorithm training, updating, or validation are insufficiently communicated or lack explicit consent and formal authorization.
Uncertainty, accountability, and human control	Undermined patient's ability to understand who is responsible for clinical decisions and how errors or harms related to AI-supported recommendations are addressed.

(*Continued*)

Table 1 (Continued).

Ethical Issue	Relevance to Informed Consent
Automation bias and erosion of shared decision-making	Undermined patient's ability to participate meaningfully in shared decision-making when AI outputs are perceived as authoritative or when systems do not integrate patients' values and preferences.
AI-mediated communication and patient education tools (e.g., chatbots)	Undermined patient's ability to participate actively in IC when AI-mediated communication provides information without adequate personalization, contextualization, or integration into clinician–patient relationship and related discussions.
Trust, acceptance, and legitimacy of AI-supported care	Undermined patient's ability to trust clinical recommendations when AI use is perceived as opaque, imposed, or misaligned with professional judgment and patient values.
Design without end-user involvement (lack of co-development)	Undermined patient's ability to see their values, preferences, and lived experiences reflected in AI-supported decisions when systems are developed without patient or clinician input.

Abbreviations: AI, artificial intelligence; IC, informed consent.

A common objection is that clinicians' reasoning is also partly opaque. The proverbial "clinical eye" integrates experience, knowledge, and intuition in ways that are often difficult to explain or reproduce. This argument has been used to downplay the black box dilemma and to prioritize correctness, safety, usefulness, and validation of outputs, over explanation of reasoning, regardless of human or chatbot origin.[41] Nonetheless, we argue that this position implicitly assumes that human and chatbot reasoning are interchangeable in how they influence behavior and subsequent reasoning. This assumption is not necessarily justified, as illustrated below through a narrative example.

In Kazuo Ishiguro's novel "Klara and the Sun", Klara is a robot companion who supports Josie, a girl with a life-threatening disease.[42] Although perceived as highly intelligent and capable of sophisticated reasoning, Klara comes to believe that communicating with the sun could lead to Josie's cure. By abductive reasoning (eg., inferring the most plausible explanation from observed patterns),[43] most humans would consider such a causal relationship scientifically implausible. Human reasoning, despite its partial opacity, is grounded in broadly shared evolutionary and cognitive frameworks. No comparable alignment can be assumed for machines, including chatbots. Thus, even when chatbot outputs appear correct and scientifically sound, they may arise from associations that are epistemically misaligned with human expectations. In complex settings such as healthcare, such misalignments could potentially influence decision-making if unrecognized. While fictional, this example illustrates a key conceptual point: when interacting with chatbots, responses appear fluent, authoritative and coherent, which may obscure the possibility that their underlying reasoning differs substantially from human reasoning and may include weakly grounded or implausible associations. Importantly, in this context, *shared reasoning fragility* should be clearly distinguished from related concepts. First, it differs from automation bias, which describes overreliance on automated systems; *shared reasoning fragility* may arise even when clinicians remain actively engaged and do not defer decisions. Second, it is distinct from hallucinations, which refer to factually incorrect outputs; fragility may occur even when outputs are factually correct but derived through reasoning pathways that do not align with clinician expectations. Third, while related to epistemic opacity, the inaccessibility of the internal processes underlying model outputs, *shared reasoning fragility* does not primarily concern the opacity itself, but rather how such opacity interacts with human reasoning during ongoing dialogue. In this sense, epistemic opacity can be understood as a precondition, whereas *shared reasoning fragility* denotes a potential consequence of the interaction between opaque machine processes and human cognition.

Finally, A second key distinction concerns the difference between chatbot-based systems and traditional classification-based clinical decision support systems (CDSSs). In conventional CDSSs settings, clinicians typically complete most reasoning before consulting the system at a final decision point. By contrast, chatbots may be used throughout the entire reasoning process, from initial hypothesis generation to diagnostic work-up, interpretation of results, and treatment selection. This enables a form of continuous co-reasoning, in which clinicians are repeatedly exposed to interactions between two fundamentally different modes of association and inference. This interaction is both unprecedented and

pervasive. Given the novelty of this interaction and the limited experience in managing it, we define *shared reasoning fragility* as the potential instability arising from continuous co-reasoning and from the current limited ability to anticipate and manage its effects. In this context, "fragility" does not imply an inherently negative phenomenon, but rather highlights the need for systematic investigation to understand and appropriately manage these interactions.

In the following section, we present descriptive examples from the literature suggesting the possible presence of *shared reasoning fragility* in chatbot-assisted antibiotic prescribing. These examples may inform the design of empirical studies and support conceptual extension to other areas of healthcare.

Shared Reasoning Fragility in Antibiotic Prescribing

Several studies have investigated the use of general-purpose LLM-based chatbots as CDSSs for antibiotic prescribing, primarily focusing on performance metrics such as accuracy and the presence of incorrect or potentially harmful responses.[5,6,9–15] These performance measures are not the focus of this work, although we share the impression that technical advances, ideally supported by interaction with medical professionals during development, are likely to enhance performance. Instead, we consider selected illustrative examples concerning the nature and potential cognitive impact of suboptimal responses within the framework of *shared reasoning fragility.* Importantly, these examples are not intended as empirical evidence but as conceptual prompts suggesting that such fragility may exist and warrants systematic investigation to confirm or refute it. No quantitative analysis was performed in the present theoretical paper, and dedicated empirical studies are required before drawing any clinical implications.

Among the largest studies on the topic, De Vito et al assessed the performance of 14 LLM-based tools for recommending antibiotic choice, dosage, and duration using simulated infectious diseases vignettes.[15] While methodologically robust, the study provides limited insight into *shared reasoning fragility.* The models were queried at a late stage of the prescribing process, with diagnosis, infection site, causative agent, and susceptibility profile already defined. In this respect, the scenario resembles traditional CDSSs, where systems are applied to a final classification task after most clinical reasoning has been completed. Although some variability and misalignment with guidelines were observed, the study reported only on structured outputs (eg., drug, dose, duration) rather than full natural language outputs. This limits the ability to examine how chatbot responses might interact with clinician reasoning, particularly how arguments are framed and how convincingly suboptimal recommendations are presented.

Smaller studies provide more informative qualitative insights into shared reasoning. For example, Maillard et al evaluated ChatGPT 4 by requesting comprehensive management plans from a positive blood culture, including work-up, antibiotic therapy, source control, and follow-up.[11] The chatbot generated detailed and well-structured responses across 44 anonymised real cases. Only one full conversation was reported, involving methicillin-susceptible *Staphylococcus aureus* (MSSA) bloodstream infection (BSI). Nonetheless, some illustrative aspects can be noted from this single example. The clinical presentation included features compatible with pneumonia (polypnea and right basal crepitations), but also inconsistent elements (absence of dyspnoea and cough), making pneumonia uncertain. Clinical features were also compatible with skin and soft tissue infection (erythematous plaque at an ulcerated forearm lesion). The chatbot appropriately identified at least one of these potential sources of BSI. Within its management plan, it recommended continuing amoxicillin–clavulanate (initiated empirically for suspected pneumonia) based on its "good coverage for skin and lung infections" and proposed a reasonable diagnostic work-up (eg., exclusion of endocarditis) with clinical and laboratory monitoring. While this reasoning may appear plausible, amoxicillin/clavulanate is not an optimal choice for treating MSSA BSI.[44,45] For infectious diseases specialists, this limitation is evident; however, non-specialists - who may benefit most from CDSS support - may be more susceptible to such framing based on shared reasoning. Unlike traditional CDSSs, the chatbot provides fluent, contextually coherent justifications that, during shared reasoning, may subtly shift clinicians' attention away from critical evaluative questions, such as the appropriateness of the antibiotic for the specific pathogen.

Notably, similar risks may arise even from correct elements of a response. For example, guideline-concordant diagnostic suggestions may convey a sense of completeness when expressed in natural language aligned with clinicians' reasoning. In the same example, pneumonia was suspected but not confirmed, yet a chest X-ray was not included in the proposed diagnostic work-up, and this omission was not highlighted in the study.[11] This illustrates how gaps may be less salient within otherwise coherent responses. A comparable observation emerged from our own interaction with ChatGPT

4o in a fictional case of a 70-year-old man with ventilator-associated pneumonia and septic shock in a ward endemic for *Klebsiella pneumoniae* carbapenemase (KPC)-producing *K. pneumoniae*.[46] Although the proposed plan broadly aligned with guidelines, a basic safety check (eg., assessment of antibiotic allergies) was initially overlooked. Notably, this omission was not immediately apparent, despite its routine importance in standard clinical reasoning. At a conceptual level, we hypothesize that this may reflect *shared reasoning fragility*, whereby interaction with chatbot outputs subtly reshapes the structure of clinical reasoning.

These observations remain illustrative and do not support definitive conclusions. However, they suggest the plausibility of *shared reasoning fragility* and the need for targeted empirical research to determine its presence, extent, and clinical implications. Our concept partially resonates with "epistemia", as recently described by Quattrociocchi et al, whereby linguistic plausibility becomes a structural substitute for epistemic evaluation.[47] In our view, while different, the two concepts are complementary: *epistemia* describes the epistemic substitution mechanism, whereas *shared reasoning fragility* focuses on the interactional conditions under which such mechanisms may influence clinical reasoning.

Finally, traditional clinical reasoning typically involves iterative selection of a manageable subset of hypotheses to optimize patient outcomes.[48,49] From this perspective, sharing reasoning with chatbots may either improve or undermine patient benefit by subtly shifting hypothesis selection throughout decision-making phases (eg., through non-consideration of relevant tests or safety checks). In our view, this further underscores the need for systematic investigation of shared reasoning processes and for strategies to mitigate potential fragility, alongside broader efforts to address risks in human-AI interaction, including amplification of cognitive biases or the influence of implicit value prioritization in AI-generated responses.[50–52]

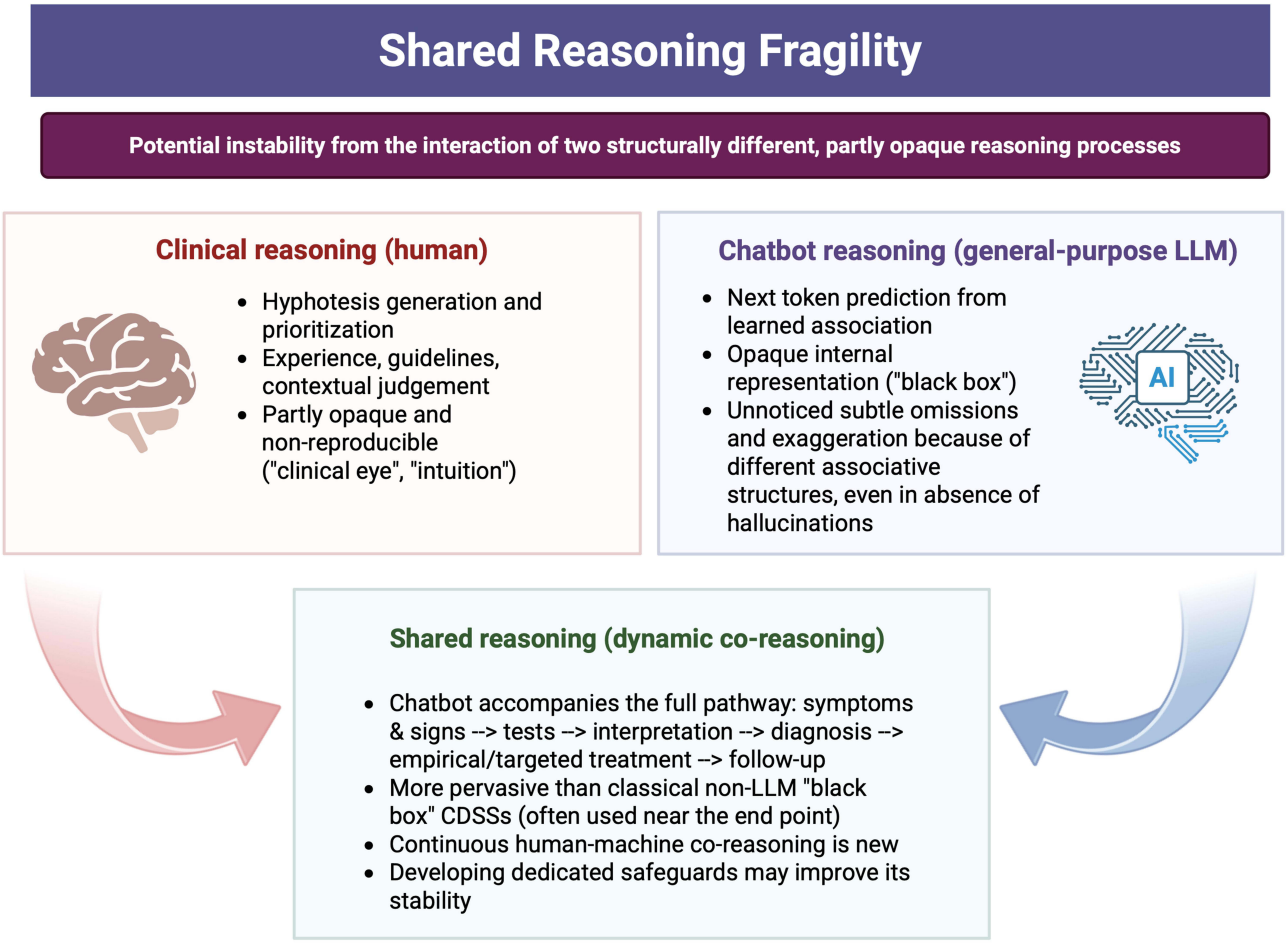

Figure 2 Theoretical framework for shared reasoning fragility through the example of antibiotic prescribing. Created in BioRender. Giacobbe, D.R. (2026) https://BioRender.com/wofjwey.
Abbreviations: AI, artificial intelligence; CDSSs, clinical decision support systems; LLM, large language model.

Conclusion

In this article, we provided a theoretical framework for *shared reasoning fragility*, illustrated through examples related to antibiotic prescribing (Figure 2). These examples are strictly illustrative and do not constitute empirical evidence. These baseline, theoretical considerations are intended to inform future quantitative and qualitative research, including vignette-based experiments, think-aloud studies, simulation studies, and assessments of whether chatbot interaction affects diagnostic completeness, hypothesis generation, or omission rates. Further research should also examine how *shared reasoning fragility* may manifest in clinical reasoning, supervision, trust calibration, and moral experience. More broadly, our framework highlights the need to move beyond accuracy-based evaluations of chatbot outputs toward a more comprehensive understanding of how these systems interact with, and potentially reshape, clinical reasoning processes. It may also serve to identify mitigating strategies, such as structured prompting, human oversight, antimicrobial stewardship workflows, and institutional safeguards. Clarifying these dimensions will be essential for the safe, effective, and ethically grounded integration of chatbots into clinical practice.

Acknowledgments

After the authors wrote the manuscript without assistance from generative AI, a LLM-based chatbot (ChatGPT 5.2) was employed to enhance its readability. Subsequently, the final text was thoroughly reviewed by all authors to ensure the preservation of content and meaning.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work was not funded.

Disclosure

Outside the submitted work, Matteo Bassetti has received funding for scientific advisory boards, travel, and speaker honoraria from Cidara, Gilead, Menarini, MSD, Mundipharma, Pfizer, and Shionogi. Outside the submitted work, Daniele Roberto Giacobbe reports investigator-initiated grants from Pfizer, Shionogi, Menarini, Tillotts Pharma, bioMérieux, Advanz Pharma, Tillotts Pharma, and Gilead Italia, travel support from Pfizer, bioMérieux, and Shionogi, and speaker/advisor fees from Pfizer, bioMérieux, Shionogi, Advanz Pharma, Menarini, MSD, and Tillotts Pharma. The authors report no other conflicts of interest.

References

1. OpenAI. Introducing ChatGPT. Available from: https://openai.com/index/chatgpt/. Accessed December 6, 2025.
2. Chatterji A, Cunningham T, Deming DJ, et al. How people use ChatGPT. *NBER Working Paper*. 2025;34255. doi:10.3386/w34255
3. Google. A new era of intelligence with Gemini 3. Available from: https://blog.google/products/gemini/gemini-3/#note-from-ceo. Accessed January 7, 2026.
4. Secondtalent. Claude AI statistics and user trends for 2026. Available from: https://www.secondtalent.com/resources/claude-ai-statistics. Accessed January 7, 2026.
5. De Vito A, Geremia N, Marino A, et al. Assessing ChatGPT's theoretical knowledge and prescriptive accuracy in bacterial infections: a comparative study with infectious diseases residents and specialists. *Infection*. 2024. doi:10.1007/s15010-024-02350-6
6. Fisch U, Kliem P, Grzonka P, Sutter R. Performance of large language models on advocating the management of meningitis: a comparative qualitative study. *BMJ Health Care Inform*. 2024;31(1):e100978. doi:10.1136/bmjhci-2023-100978
7. Giacobbe DR, Marelli C, La Manna B, et al. Advantages and limitations of large language models for antibiotic prescribing and antimicrobial stewardship. *NPJ Antimicrob Resist*. 2025;3(1):14. doi:10.1038/s44259-025-00084-5
8. Goodman KE, Tamma PD. Large language models for antibiotic prescribing-moving the needle from 'parlour trick' to practical tool. *Clin Microbiol Infect*. 2025;31(8):1260–1262. doi:10.1016/j.cmi.2025.05.014
9. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? *Lancet Infect Dis*. 2023;23(4):405–406. doi:10.1016/S1473-3099(23)00113-5

10. Lorenzoni G, Garbin A, Brigiari G, Papappicco CAM, Manfrin V, Gregori D. Large language models in action: supporting clinical evaluation in an infectious disease unit. *Healthcare*. 2025;13(8):879. doi:10.3390/healthcare13080879
11. Maillard A, Micheli G, Lefevre L, et al. Can chatbot artificial intelligence replace infectious diseases physicians in the management of bloodstream infections? A prospective cohort study. *Clin Infect Dis*. 2024;78(4):825–832. doi:10.1093/cid/ciad632
12. Montiel-Romero S, Rajme-Lopez S, Roman-Montes CM, et al. Recommended antibiotic treatment agreement between infectious diseases specialists and ChatGPT((R)). *BMC Infect Dis*. 2025;25(1):38. doi:10.1186/s12879-024-10426-9
13. Ngoc Nguyen O, Amin D, Bennett J, et al. GP or ChatGPT? Ability of large language models (LLMs) to support general practitioners when prescribing antibiotics. *J Antimicrob Chemother*. 2025;80(5):1324–1330. doi:10.1093/jac/dkaf077
14. Tao H, Liu L, Cui J, Wang K, Peng L, Nahata MC. Potential use of ChatGPT for the treatment of infectious diseases in vulnerable populations. *Ann Biomed Eng*. 2024;52(12):3141–3144. doi:10.1007/s10439-024-03600-2
15. De Vito A, Geremia N, Bavaro DF, et al. Comparing large language models for antibiotic prescribing in different clinical scenarios: which performs better? *Clin Microbiol Infect*. 2025;19. doi:10.1016/j.cmi.2025.03.002
16. Giacobbe DR, Bassetti M. Beyond explainability: introducing shared reasoning fragility. *Int J Med Inform*. 2025;207:106188. doi:10.1016/j.ijmedinf.2025.106188
17. Giacobbe DR, Marelli C, Muccio M, et al. From Klebsiella and Candida to artificial intelligence: a perspective from infectious diseases doctors and researchers. *Front Med*. 2025;12:1676920. doi:10.3389/fmed.2025.1676920
18. Xu H, Shuttleworth KMJ. Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". *Intelligent Medicine*. 2024;4(1):52–57.
19. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. *iScience*. 2024;27(5):109713. doi:10.1016/j.isci.2024.109713
20. Nassiri K, Akhloufi MA. Recent advances in large language models for healthcare. *BioMedInformatics*. 2024;4(2):1097–1143. doi:10.3390/biomedinformatics4020062
21. Schwartz IS, Link KE, Daneshjou R, Cortes-Penfield N. Black box warning: large language models and the future of infectious diseases consultation. *Clin Infect Dis*. 2024;78(4):860–866. doi:10.1093/cid/ciad633
22. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023;330(9):866–869. doi:10.1001/jama.2023.14217
23. Ali S, Akhlaq F, Imran AS, Kastrati Z, Daudpota SM, Moosa M. The enlightening role of explainable artificial intelligence in medical & healthcare domains: a systematic literature review. *Comput Biol Med*. 2023;166:107555.
24. Amann J, Vetter D, Blomberg SN, et al. To explain or not to explain?-Artificial intelligence explainability in clinical decision support systems. *PLOS Digit Health*. 2022;1(2):e0000016. doi:10.1371/journal.pdig.0000016
25. Giacobbe DR, Marelli C, Guastavino S, et al. Explainable and interpretable machine learning for antimicrobial stewardship: opportunities and challenges. *Clin Ther*. 2024;46(6):474–480. doi:10.1016/j.clinthera.2024.02.010
26. Renftle M, Trittenbach H, Poznic M, Heil R. What do algorithms explain? The issue of the goals and capabilities of Explainable Artificial Intelligence (XAI). *Humanit Soc Sci Commun*. 2024;11(1):760. doi:10.1057/s41599-024-03277-x
27. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Mach Intell*. 2019;1(5):206–215. doi:10.1038/s42256-019-0048-x
28. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. 2021;3(11):e745–e750. doi:10.1016/S2589-7500(21)00208-9
29. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, et al. Application of artificial intelligence in shared decision making: scoping review. *JMIR Med Inform*. 2022;10(8):e36199. doi:10.2196/36199
30. Busch F, Adams LC, Bressem KK. Biomedical ethical aspects towards the implementation of Artificial Intelligence in Medical Education. *Med Sci Educ*. 2023;33(4):1007–1012. doi:10.1007/s40670-023-01815-x
31. Cesaro A, Hoffman SC, Das P, de La Fuente-Nunez C. Challenges and applications of artificial intelligence in infectious diseases and antimicrobial resistance. *NPJ Antimicrob Resist*. 2025;3(1):2. doi:10.1038/s44259-024-00068-x
32. Naik N, Hameed BMZ, Shetty DK, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg*. 2022;9:862322. doi:10.3389/fsurg.2022.862322
33. Nogaroli R, Faleiros Júnior JLDM. Ethical challenges of artificial intelligence in medicine and the triple semantic dimensions of algorithmic opacity with its repercussions to patient consent and medical liability. In: Sousa Antunes H, Freitas PM, Oliveira AL, Martins Pereira C, de Sequeira E V, Barreto Xavier L, editors. *Multidisciplinary Perspectives on Artificial Intelligence and the Law*. Springer International Publishing; 2024:229–248.
34. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med*. 2018;15(11):e1002689. doi:10.1371/journal.pmed.1002689
35. Giacobbe DR, Grossi AA, Bassetti M, de La Fuente-Nunez C. The future of antibiotics and artificial intelligence: some thoughts from discovery to bedside. *Infect Dis Ther*. 2025. doi:10.1007/s40121-025-01288-y
36. Jiang Y, Chen J, Yang D, et al. CoMT: Chain-of-medical-thought reduces hallucination in medical report generation. 2025:1–5.
37. Wei J, Wang X, Schuurmans D, et al. Chain of thought prompting elicits reasoning in large language models. *ArXiv*. 2022. abs/2201.11903.
38. Yang H, Zhao Q, Li L. How chain-of-thought works? Tracing information flow from decoding, projection, and activation. *ArXiv*. 2025;abs/2507.20758.
39. Tosato T, Helbling S, Ramos YJM, et al. Persistent instability in LLM's personality measurements: effects of scale, reasoning, and conversation history. *ArXiv*. 2025. abs/2508.04826.
40. Korbak T, Balesni M, Barnes E-B, et al. Chain of thought monitorability: a new and fragile opportunity for AI safety. *ArXiv*. 2025. abs/2507.11473.
41. Hanscheid T, Carrasco J, Grobusch MP. Re: 'Comparing large language models for antibiotic prescribing in different clinical scenarios' by De Vito et al. *Clin Microbiol Infect*. 2025;31(12):2102–2103. doi:10.1016/j.cmi.2025.08.006
42. Ishiguro K. *Klara and the Sun*. Faber and Faber; 2021.
43. Sandoval-Hernández A, Rutkowski DJ. Embracing complexity: abductive reasoning as a versatile tool for analyzing international large-scale assessments. *Educ Assessment Evaluat AccounT*. 2025;37(2):255–271. doi:10.1007/s11092-024-09449-2

44. Giacobbe DR, Dettori S, Corcione S, et al. Emerging treatment options for acute bacterial skin and skin structure infections and bloodstream infections caused by staphylococcus aureus: a comprehensive review of the evidence. *Infect Drug Resist*. 2022;15:2137–2157. doi:10.2147/IDR.S318322
45. Paul M, Zemer-Wassercug N, Talker O, et al. Are all beta-lactams similarly effective in the treatment of methicillin-sensitive Staphylococcus aureus bacteraemia? *Clin Microbiol Infect*. 2011;17(10):1581–1586. doi:10.1111/j.1469-0691.2010.03425.x
46. Giacobbe DR, Sanguinetti M, Bassetti M. Re: 'Comparing large language models for antibiotic prescribing in different clinical scenarios' by De Vito et al. *Clin Microbiol Infect*. 2025;31(8):1408–1409. doi:10.1016/j.cmi.2025.03.022
47. Quattrociocchi W, Capraro V, Perc M. Epistemological fault lines between human and artificial intelligence. *ArXiv*. 2025;abs/2512.19466.
48. Keeling G, Nyrup R. Explainable machine learning, patient autonomy, and clinical reasoning. In: Véliz C, editor. *Oxford Handbook of Digital Ethics*. Oxford University Press; 2023.
49. Stanley DE, Campos DG. The logic of medical diagnosis. *Perspect Biol Med*. 2013;56(2):300–315. doi:10.1353/pbm.2013.0019
50. Guiducci L, Saulle C, Dimitri GM, et al. Dialogical AI for cognitive bias mitigation in medical diagnosis. *Appl Sci*. 2026;16(2):710. doi:10.3390/app16020710
51. Ye A, Moore J, Novick R, Zhang AX. Language models as critical thinking tools: a case study of philosophers. *ArXiv*. 2024;abs/2404.04516.
52. Goldberg C, Balicer RD, Bhat M, et al. The missing dimension in clinical AI: making hidden values visible. *NEJM AI*. 2026;3(2):AIp2501266. doi:10.1056/AIp2501266