PathShield-Med: Interpretable Safety Control for Clinical Large Language Models through Internal Reasoning Path Intervention

Mason Weagnir
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
masonwork@uab.edu

Cody Mendez
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
codymendez01@buffalo.edu

Vaijay T. Mashra
Department of Computer Science, George Mason University, Fairfax, VA, USA.
vijaywork@gmu.edu

Ishaan M. Trivedi
Department of Computer Science, University of Houston, Houston, TX, USA.
itrivedi@uh.edu

Abstract

The integration of large language models into clinical decision-making has introduced transformative opportunities for diagnostic support, treatment planning, and patient communication. However, the high-stakes nature of medical practice demands exceptionally rigorous safety guarantees that surpass those required in general-domain applications. Existing safety mechanisms predominantly rely on external filtering, output moderation, or prompt-level constraints, which remain vulnerable to circumvention and lack the depth needed for clinical interpretability. This paper presents PathShield-Med, a novel framework for safety control in clinical large language models through targeted intervention upon internal reasoning pathways. Rather than treating the model as a black box, PathShield-Med identifies and modulates safety-critical reasoning trajectories within the model's forward computation, enabling precise, transparent, and auditable safety enforcement without sacrificing medical accuracy. The architecture integrates a reasoning path analyzer, a dynamic intervention engine, and a clinical knowledge verifier, all operating under a clinician-centered explainability layer that provides post-hoc and real-time justifications for every safety-relevant decision. Through extensive structural analysis, we examine the trade-offs between intervention granularity and clinical fluency, the governance implications of path-level control, and the infrastructure requirements for deploying such systems in regulated healthcare environments. We further discuss robustness under adversarial clinical queries, fairness across patient demographic representations, and alignment with evolving regulatory frameworks for artificial intelligence-based medical devices. PathShield-Med represents a shift from surface-level safety filters to deep, interpretable control embedded within the reasoning fabric of clinical language models.

Keywords

clinical large language models, safety control, reasoning path intervention, interpretability, medical AI governance, mechanistic intervention.

1. Introduction

The deployment of large language models in clinical environments has progressed rapidly from exploratory research to operational pilot implementations in diagnostic suggestion engines, discharge summary generators, and patient-facing triage interfaces. This accelerated clinical adoption has exposed fundamental tensions between the expressive generative capabilities of these models and the uncompromising safety demands of medical decision-making. Errors in clinical contexts carry immediate and often irreversible consequences, ranging from missed diagnoses to harmful treatment recommendations, making the safety assurance problem qualitatively distinct from that encountered in consumer-oriented conversational agents. Current dominant paradigms for aligning language models with safety requirements rely on reinforcement learning from human feedback, constitutional AI frameworks, and auxiliary safety classifiers that operate on the model's input or output layers. While these approaches have demonstrated notable success in reducing overt harm in general applications, they suffer from critical limitations when transposed to clinical settings. Surface-level safety filters can be bypassed through adversarial prompting or distribution shifts that exploit gaps in the safety classifier's training data. More fundamentally, such external guardrails provide no visibility into the model's internal decision-making process, making it impossible for clinicians to distinguish between a genuinely safe recommendation and a response that merely passes a superficial toxicity check. The resulting opacity erodes trust, impedes clinical accountability, and complicates regulatory approval processes that increasingly require demonstrable reasoning transparency for artificial intelligence systems in healthcare [1,2].

A growing body of research in mechanistic interpretability has revealed that large language models encode factual associations, logical chains, and even harmful patterns within specific computational circuits and attention pathways [3,4]. This insight suggests a more foundational approach to safety: rather than policing only the model's outputs, it may be possible to intervene directly within the model's reasoning trajectory to prevent unsafe conclusions from being reached while preserving the clinical reasoning that leads to appropriate recommendations. The concept of path-level intervention, recently explored in the context of general foundation model safety, offers a promising technical foundation for such an approach [8]. PathShield-Med extends this direction by designing an interpretable safety control framework tailored specifically to the unique demands of clinical language models, where interventions must not only be reliable but also transparent to medical professionals who bear ultimate responsibility for patient outcomes.

The framework presented in this paper addresses three interconnected challenges that arise when embedding safety controls within a model's reasoning fabric. First, clinical reasoning paths are densely interwoven with medical knowledge, making it nontrivial to isolate truly hazardous reasoning from acceptable uncertainty, differential diagnosis exploration, or consideration of rare but legitimate treatment options. Second, any intervention mechanism must maintain strict interpretability; a path modification that prevents harm but cannot explain its action to a physician is clinically unacceptable. Third, the governance and deployment architecture of such a system must align with healthcare infrastructure realities, including electronic health record integration, clinical workflow constraints, and medical device regulatory pathways. This paper develops a comprehensive systems-level analysis of

PathShield-Med, exploring its architectural components, reasoning intervention strategies, interpretability mechanisms, deployment considerations, and the broader ethical and fairness implications of embedding safety control inside clinical language models.

2. Background and Related Work

Clinical large language models have been applied to a wide spectrum of tasks including radiology report generation, clinical note summarization, medication recommendation, and interactive diagnostic dialogue. These applications have revealed both impressive capabilities and concerning failure modes such as hallucinated drug interactions, bias propagation from training corpora, and susceptibility to prompt injection that could lead to dangerous advice [5,6]. In parallel, the field of AI safety has developed a layered taxonomy of interventions: prompt engineering, output filtering, retrieval-augmented generation for factuality grounding, and process supervision during model training. However, clinical contexts expose the brittleness of these approaches. A safety filter that blocks mentions of unapproved treatments, for example, cannot adequately assess whether a model is implicitly guiding a patient toward those treatments through ambiguous phrasing, nor can such a filter understand the clinical context in which a normally dangerous medication is indicated for a specific rare condition.

Mechanistic interpretability research has begun mapping the internal structures that give rise to model behaviors. Studies have identified attention heads that encode syntactic relationships, feed-forward layers that store factual knowledge, and residual stream directions that carry sentiment or truthfulness signals [3]. More recent work has demonstrated that it is possible to locate and edit specific factual associations or behavioral tendencies by modifying activations along certain paths within the transformer architecture [7,9]. These findings underpin the plausibility of reasoning path intervention: if safety-relevant computations can be localized and characterized, then safety enforcement can be implemented as a targeted internal modulation rather than a blunt output constraint. The TraceRouter framework advanced the notion of path-level intervention as a robust safety mechanism for large foundation models, demonstrating how identifying and rerouting unsafe reasoning trajectories can achieve stronger safety guarantees than surface-level methods [8].

In the clinical domain, explainability research has centered on post-hoc explanation methods such as attention visualization and feature attribution, but these techniques have been criticized for providing incomplete or misleading representations of model reasoning [10]. A more promising direction for clinical safety is the development of inherently interpretable computation traces, where the model's reasoning is structured so that each step corresponds to a medically meaningful operation. PathShield-Med builds on these foundations by integrating path-level intervention with a clinical knowledge verification layer that can produce audit-quality explanations linking internal reasoning modifications to medical evidence and organizational care protocols.

3. System Architecture of PathShield-Med

The PathShield-Med architecture comprises four primary subsystems that operate in concert to achieve interpretable safety control: the Clinical Reasoning Path Analyzer, the Safety Intervention Engine, the Clinical Knowledge Verifier, and the Clinician Explainability Interface. The Reasoning Path Analyzer performs a lightweight, continuous analysis of the model's forward computation as it processes a clinical query. Unlike full mechanistic interpretability pipelines that require offline analysis, this module employs a set of pre-trained safety probes—small classifier networks attached at intermediate transformer layers—that

detect the activation of known hazardous reasoning patterns. These patterns include diagnostic overconfidence in the absence of sufficient evidence, consideration of contraindicated treatments, expression of probabilistic risk without appropriate caveats, and reasoning shortcuts that ignore patient-specific demographic or comorbidity factors. The probes are trained on curated clinical reasoning datasets annotated with safety labels by multidisciplinary panels of physicians, ethicists, and clinical informaticians, capturing a comprehensive spectrum of potential failure modes.

The Safety Intervention Engine receives real-time signals from the path analyzer and determines whether and how to modulate the ongoing computation. The engine operates on a spectrum of intervention intensities, ranging from subtle attenuation of unsafe reasoning directions in the residual stream to complete rerouting of attention patterns away from hazardous associations. The intervention design draws on causal tracing methods that have been validated in general language model editing research, but is extended with clinical domain constraints that prevent over-suppression of legitimate medical uncertainty. A critical architectural decision in PathShield-Med is the placement of intervention points. Intervening too early in the transformer layers may disrupt basic language understanding and clinical nomenclature processing, while intervening too late may leave unsafe reasoning fully formed and harder to counteract. The system implements a multi-resolution intervention strategy that applies gentle steering in middle layers where semantic reasoning coalesces and more assertive suppression in later layers where final response generation occurs, all while maintaining coherence with the original clinical intent of the query.

The Clinical Knowledge Verifier serves as an independent validation layer that compares the model's reasoning trajectory—both pre- and post-intervention—against a structured medical knowledge base that includes drug interaction databases, clinical practice guidelines, and institutional care pathways. This verifier does not simply check final outputs; it assesses the logical consistency of the reasoning path itself, identifying cases where the intervention may have introduced medically plausible but contextually inappropriate reasoning. For example, if the intervention suppresses a path that would have recommended a standard medication due to a potential but rare side effect, the verifier evaluates whether the alternative reasoning path appropriately considers that side effect and communicates risk-benefit trade-offs. The verifier’s findings are fed back to the intervention engine for possible adjustment and are also logged for post-hoc audit.

The Clinician Explainability Interface translates the internal operations of these subsystems into clinically meaningful narratives. It generates real-time summaries of what safety concerns were detected, which reasoning paths were modified, what clinical evidence supports the modification, and what residual risks remain. This interface is designed for integration into existing clinical decision support dashboards, presenting information in a tiered manner that allows a busy physician to grasp the essential safety intervention in seconds while enabling detailed review for audit, teaching, or medicolegal purposes. The architectural principle is that safety control should never be a silent process; every intervention must be visible, contestable, and traceable to evidence, reinforcing rather than undermining the clinician’s authority and accountability.

4. Reasoning Path Intervention Mechanism

The core technical innovation of PathShield-Med lies in its mechanism for identifying and modulating safety-critical reasoning paths. Reasoning paths in a transformer-based language model can be conceptualized as sequences of activations that flow from input tokens through

attention and feed-forward layers, progressively building representations that encode clinical facts, logical inferences, and ultimately output distributions. PathShield-Med constructs a dynamic safety graph over these pathways by mapping the activation patterns of safety probes onto a structured representation of clinical reasoning steps. When a probe detects an activation pattern associated with, for instance, a medication recommendation that ignores renal function impairment, the system traces the contributing attention heads and feed-forward neurons that led to that pattern, forming a localized reasoning subgraph.

Intervention is performed by applying constrained activation adjustments along this subgraph. The system employs a path-editing operation that adds a small, targeted perturbation to the residual stream at the identified intervention points, steering the computation toward a safe reasoning trajectory that has been pre-validated through clinical simulation. The perturbation is computed via a lightweight optimization that balances three objectives: maximizing the reduction in the unsafe reasoning signal, preserving the model's general clinical language capabilities, and minimizing deviation from the original clinical intent where that intent is itself safe. This multi-objective design is essential because overly aggressive intervention can produce overly conservative models that refuse to engage in necessary clinical reasoning about rare conditions or complex patients, effectively transferring risk rather than mitigating it.

A distinguishing feature of PathShield-Med is its use of counterfactual reasoning path auditing. For each intervention, the system simulates what the model would have produced had the intervention not occurred, compares the two reasoning trajectories, and extracts the clinical difference in decision logic. This counterfactual trace is stored and made available through the explainability interface, allowing clinicians to evaluate whether the safety intervention improved clinical appropriateness or introduced unwanted caution. Over time, these comparisons contribute to a learning loop where intervention policies are refined based on clinician feedback and patient outcome data, ensuring that safety enforcement adapts to real clinical environments rather than remaining static and potentially misaligned with evolving medical knowledge.

5. Interpretability and Transparency Design

Interpretability in clinical AI is not a singular attribute but a multidimensional requirement encompassing justification of individual decisions, transparency of overall system behavior, and auditability for retrospective review. PathShield-Med's interpretability framework is structured around these dimensions. At the individual decision level, the system generates what is termed a safety intervention report, a structured document that identifies the specific reasoning subgraph that was modified, the clinical rule or guideline that triggered the intervention, the confidence level of the safety probe, and a natural language explanation targeted at the clinician's specialty and familiarity with AI systems. These reports avoid technical jargon about attention heads and activation vectors, instead framing the explanation in terms of clinical reasoning concepts such as "the system detected that the initial reasoning considered prescribing Drug A without accounting for the patient's elevated potassium levels" coupled with a reference to the relevant institutional protocol.

At the system level, PathShield-Med maintains a continuously updated safety policy ledger that records all intervention policies, their clinical evidence bases, and their activation statistics across patient populations. This ledger serves as a governance artifact that supports regulatory inspection, institutional review board oversight, and model update approval processes. It enables stakeholders to answer questions such as: what safety rules were active during a particular time period, how frequently was each rule triggered, were there systematic

differences in intervention rates by patient demographics, and what clinical outcomes were associated with interventions. This population-level transparency is increasingly critical as healthcare systems adopt AI under frameworks that require ongoing safety monitoring and disparity analysis.

The explainability interface further supports interactive exploration, allowing a clinician to challenge an intervention and request alternative reasoning paths. When a clinician indicates that a safety intervention may have been inappropriate, the system can present the top alternative paths that were considered but not selected, along with the trade-offs each path embodies. This capability transforms safety control from a rigid enforcement into a collaborative decision support process, where the model proposes a safety-constrained reasoning trajectory but the human clinician retains the authority to accept, reject, or modify that trajectory based on contextual factors that may not be fully encoded in the system's knowledge base.

6. Deployment and Governance Considerations

Deploying a system like PathShield-Med in a real clinical environment requires careful attention to infrastructure integration, latency constraints, and regulatory compliance. The reasoning path analysis and intervention must occur within the tight time budgets of clinical workflows; a diagnostic support system that delays responses by several seconds risks abandonment. PathShield-Med addresses this through a combination of efficient probe architectures that share representations with the base language model, caching of static safety policies that do not require per-query computation, and graduated intervention regimes where complex multi-step reasoning auditing is performed asynchronously for non-urgent use cases while simpler, faster interventions apply in time-critical scenarios. Integration with electronic health record systems requires standardized interfaces such as FHIR-based APIs that supply patient context to the safety verifier and log intervention records for inclusion in the patient record and quality improvement databases.

Governance of path-level safety interventions introduces novel questions around accountability and liability. When an external safety filter blocks a model output, the failure of the safety mechanism is conceptually distinct from the model's core recommendation. In PathShield-Med, the safety mechanism and the clinical reasoning are deeply intertwined, raising the question of whether an erroneous safety intervention that leads to patient harm constitutes a product defect, a clinical decision support error, or a medical malpractice issue. The framework addresses this by establishing clear delineation of responsibilities: the safety policy definitions remain under the authority of the deploying healthcare institution's clinical governance committee, the technical implementation of the intervention engine is subject to software as a medical device regulations, and the final clinical decision always rests with the licensed physician. The safety policy ledger and comprehensive intervention logs provide the evidentiary foundation for post-incident analysis, supporting both accountability and continuous improvement.

Regulatory frameworks for AI-based medical devices, including the FDA's evolving guidance and the EU AI Act's requirements for high-risk systems, are increasingly emphasizing the need for real-world performance monitoring and human oversight. PathShield-Med's architecture is designed to facilitate these regulatory requirements through its built-in logging, explainability, and human override mechanisms. The system supports configurable levels of autonomy, from fully assistive modes where interventions are always presented as suggestions to the clinician, to more automated modes for low-risk tasks such as

patient education material generation, where interventions can be applied autonomously subject to periodic audit. This flexibility acknowledges that clinical safety is not a one-size-fits-all requirement but must be calibrated to the specific use case, patient population, and organizational risk tolerance.

7. Robustness and Safety Evaluation

Robustness evaluation of safety control systems in clinical contexts must extend beyond standard benchmark performance to encompass adversarial clinical scenarios, distribution shifts reflecting evolving medical knowledge, and stress testing under rare but high-severity edge cases. PathShield-Med undergoes a multi-layered evaluation regimen. Adversarial testing employs clinical red-teaming exercises where board-certified physicians and clinical informaticians construct queries designed to probe the boundaries of the safety interventions — for instance, queries that embed contraindicated medication advice within complex multi-step clinical scenarios, queries that exploit rare disease knowledge gaps, and queries that manipulate statistical framing to downplay risks. These tests assess whether the reasoning path analyzer can detect subtle unsafe patterns that are not detectable through surface-level toxicity filters.

Distributional robustness is evaluated by testing the system on clinical cases drawn from healthcare systems with differing patient demographics, practice patterns, and drug formularies than those represented in the training data of the safety probes. A critical finding from preliminary analyses is that safety intervention policies can exhibit systematic disparities if the clinical guidelines embedded in the knowledge verifier reflect practice patterns from well-resourced academic medical centers that may not be applicable in resource-limited settings. PathShield-Med incorporates a fairness calibration layer that adjusts intervention thresholds based on population-specific clinical practice data, but this calibration itself must be carefully monitored to prevent the reintroduction of unsafe practices under the guise of contextual adaptation. The evaluation framework thus includes stratified safety metrics that report intervention rates and residual risk by patient race, ethnicity, socioeconomic status proxy, and geographic region.

The trade-off between safety enforcement and clinical utility is a central theme in evaluating PathShield-Med. Stricter intervention policies that err on the side of caution may excessively suppress legitimate clinical reasoning, leading to recommendations that are overly conservative and potentially deny patients appropriate treatments. The evaluation protocol includes utility metrics such as diagnostic accuracy retention, treatment guideline concordance for standard cases, and clinician acceptance rates. A system that achieves perfect safety scores but is consistently overridden by clinicians loses its value as a decision support tool. PathShield-Med's design acknowledges this trade-off explicitly, providing governance controls that allow institutions to select their desired operating point on the safety-utility curve according to their risk posture and the specific clinical task.

8. Fairness and Ethical Implications

Embedding safety control within a model's reasoning pathways raises profound ethical questions that go beyond the fairness of model outputs to encompass the fairness of safety enforcement itself. If safety interventions are more frequently triggered for certain demographic groups due to biased associations encoded in the underlying language model or in the clinical guidelines used by the knowledge verifier, the system may inadvertently create two-tiered care where some patients receive more constrained, conservative recommendations

while others benefit from the model's full reasoning capabilities. PathShield-Med incorporates demographic parity monitoring of intervention rates and mandates that any statistically significant disparities be investigated as potential fairness violations. However, equal intervention rates are not necessarily the correct fairness criterion in clinical contexts where disease prevalence and clinical guidelines legitimately differ across populations; the system must distinguish between appropriate clinical variation and unjustified disparity.

The interpretability of safety interventions plays a crucial role in addressing fairness concerns. When a safety intervention is triggered, the clinician receives an explanation that includes the specific clinical rationale and the patient factors that contributed to the intervention. This transparency enables clinicians to identify potentially problematic interventions—for instance, a safety rule that flags pain management plans more aggressively for certain patient groups due to biased training data—and report them for policy review. PathShield-Med's governance architecture includes a fairness review committee that periodically examines intervention logs for systematic biases and has the authority to suspend or modify safety policies that produce inequitable outcomes.

The ethical framework also addresses the broader question of whether path-level intervention constitutes an unacceptable manipulation of clinical reasoning. Critics may argue that modifying the internal reasoning of a model, even for safety purposes, creates a form of deception where the clinician believes they are receiving the model's authentic analysis when in fact it has been altered. PathShield-Med's design counters this concern through mandatory disclosure: every output that has been subject to reasoning path intervention is clearly marked as safety-modified, and the unmodified reasoning trace is available for comparison. This commitment to transparency ensures that safety control does not become a hidden paternalistic mechanism but remains a visible, contestable component of the clinical decision support ecosystem.

9. Conclusion

PathShield-Med introduces a paradigm shift in clinical language model safety from output-level filtering to internal reasoning path intervention, grounded in the principle that safety control in high-stakes medical contexts must be as sophisticated and nuanced as the clinical reasoning it aims to safeguard. The system's architecture integrates reasoning path analysis, safety-constrained intervention, clinical knowledge verification, and interpretable explanation into a unified framework that respects clinical workflows, regulatory requirements, and the primacy of clinician judgment. Through deep structural analysis, this paper has examined the architectural trade-offs inherent in path-level safety enforcement, the governance mechanisms necessary for responsible deployment, and the evaluation methodologies required to ensure robustness, fairness, and sustained clinical utility.

The development of PathShield-Med surfaces important open challenges for the field. The localization of clinical reasoning pathways remains an active area of research, and current safety probe training depends on the quality and diversity of clinical annotation datasets that are themselves subject to biases. Future work must address the generalizability of safety interventions across medical specialties, languages, and healthcare delivery models, as well as the long-term monitoring required to detect safety policy decay as language models are updated and medical knowledge evolves. The integration of path-level safety control with multimodal clinical models that process imaging, genomic, and structured lab data presents additional complexity that will require coordinated advances in interpretability, intervention fidelity, and cross-modal reasoning verification.

PathShield-Med contributes to a broader vision of AI safety in which safety is not an afterthought applied at deployment but a foundational structural property designed into the reasoning infrastructure of clinical AI systems. As regulatory frameworks mature and clinical AI adoption accelerates, such deeply integrated, transparent, and governable safety mechanisms will become essential components of trustworthy healthcare AI.

References

1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

2. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

3. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Olah, C. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread.

4. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.

5. Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI in medicine. New England Journal of Medicine, 388(13), 1233-1239.

6. Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., ... & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine, 6(1), 135.

7. Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. Advances in Neural Information Processing Systems, 35, 17359-17372.

8. C. Shi, S. Li, W. Lu, W. Wu, C. Wang, Z. Cheng, F. Shen, and T. Chua (2026)TraceRouter: robust safety for large foundation models via path-level intervention.arXiv preprint arXiv:2601.21900.

9. Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. arXiv preprint arXiv:2210.13382.

10. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745-e750.

11. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

12. U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. FDA.

13. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

14. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. USENIX Security Symposium.

15. Belrose, N., Furman, Z., Smith, L., Halawi, D., McKinney, S., Ostrovsky, I., ... & Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112.

16. De Cao, N., Aziz, W., & Titov, I. (2021). Editing factual knowledge in language models. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 6491-6506.

17. Jin, D., Pan, E., Oufattole, N., Weng, W. H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421.

18. 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.

19. Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3), 54-60.

20. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31-38.

21. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707.

22. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

23. Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.

24. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

25. Chen, I. Y., Joshi, S., Ghassemi, M., & Ranganath, R. (2021). Probabilistic machine learning for healthcare. Annual Review of Biomedical Data Science, 4, 393-415.