Reinforcement Learning-Based Dynamic Security Alignment for Autonomous Medical Decision-Making Agents

Stanley Lowe
Department of Computer Science, George Mason University, Fairfax, VA, USA.
lowe970@gmu.edu

Abstract

The increasing integration of autonomous decision-making agents into clinical workflows introduces profound challenges at the intersection of safety, adaptability, and regulatory compliance. This paper examines a systems-oriented framework for dynamic security alignment in medical agents that leverage reinforcement learning to continuously calibrate their behavior against evolving clinical, ethical, and operational constraints. Rather than focusing on algorithmic novelty, the discussion foregrounds the structural trade-offs between rigid rule-based oversight and fluid, context-aware alignment mechanisms. The architecture couples a reinforcement learning-based meta-controller with domain-specific safety monitors, enabling online reconfiguration of decision boundaries in response to distributional shifts, adversarial perturbations, and policy updates. Through an analysis of deployment infrastructures, governance models, and sustainability pressures, the paper argues that effective alignment must be treated as a continuous socio-technical process rather than a static design property. Key challenges such as fairness auditing across heterogeneous populations, resilience against adversarial manipulation of sensor and knowledge pathways, and the maintenance of alignment across system lifecycles are examined in depth. Comparative insights from large-scale industrial control systems and autonomous vehicle safety architectures inform the proposed design principles. The paper concludes with a discussion of the policy and regulatory implications of dynamically aligned medical agents, emphasizing the need for transparent audit trails, adaptive certification frameworks, and multistakeholder governance structures that can keep pace with learned behavioral updates.

Keywords

reinforcement learning, medical decision-making, AI alignment, dynamic security, autonomous agents, socio-technical systems, adversarial robustness, fairness, regulatory governance.

1. Introduction

The prospect of embedding autonomous agents into medical decision-making processes represents one of the most ambitious and high-stakes frontiers in artificial intelligence research. These agents, designed to assist or even supplant human clinicians in tasks ranging from diagnostic reasoning to treatment planning, promise to alleviate cognitive overload, reduce unwarranted clinical variation, and extend expert-level decision support to resource-constrained environments. However, the transition from supervised predictive models to autonomous decision-making agents introduces a distinct class of sociotechnical risks that cannot be addressed solely through improved predictive accuracy. Foremost among these risks is the challenge of maintaining robust and contextually appropriate behavior over time

as clinical guidelines evolve, patient populations shift, and malicious actors probe system vulnerabilities. The concept of security alignment, extending beyond traditional notions of model accuracy or static rule compliance, has therefore emerged as a critical systems requirement.

Autonomous medical agents differ from conventional clinical decision support tools in that they actively participate in the sequencing and selection of diagnostic and therapeutic actions, often under conditions of partial observability and in environments where the consequences of errors cascade across time. This temporal entanglement makes the design of safety constraints, intervention thresholds, and override mechanisms particularly complex. Static alignment approaches that encode a fixed set of ethical principles, clinical guidelines, or exclusionary rules at design time quickly become brittle when the operational context shifts. The rapid evolution of treatment protocols during the COVID-19 pandemic, for example, demonstrated how quickly static rule sets become outdated and how dangerous it can be to rely on systems that cannot dynamically renegotiate their own boundaries of acceptable action.

In response to these challenges, this paper explores a reinforcement learning-based dynamic security alignment framework. From a systems perspective, the core idea is to treat alignment not as a one-time verification procedure but as a continuous control problem in which a meta-level policy adapts the agent's safety margins, decision heuristics, and consultation protocols to the current environmental state and risk estimate. Reinforcement learning provides a natural substrate for such meta-control, given its capacity to optimize delayed outcomes in stochastic environments. By formulating safety constraints as part of the reward structure or as a learned auxiliary objective, a medical agent can, in principle, maintain alignment with a moving target of clinical acceptability. Yet the viability of such an approach depends on a host of infrastructural, institutional, and governance factors that reach far beyond algorithmic design.

This analysis deliberately shifts the center of gravity from mathematical formulation to the architectural, regulatory, and sustainability dimensions that determine whether dynamically aligned agents can be deployed safely and equitably. The discussion draws on lessons from large-scale industrial control systems, autonomous vehicle safety frameworks, and recent developments in large language model alignment to construct a holistic view of what a dynamic alignment infrastructure entails. In doing so, the paper seeks to bridge the gap between theoretical work on reinforcement learning safety and the practical realities of clinical engineering, regulatory oversight, and long-term system maintenance.

2. Background and Architectural Motivations

The evolution of reinforcement learning from game-playing benchmarks to high-consequence application domains has been accompanied by an expanding discourse on safety and alignment. Early influential work established that deep reinforcement learning agents could achieve superhuman performance in narrowly defined tasks but simultaneously exposed their brittleness to distributional shift and reward misspecification [1]. The concrete problems in AI safety enumerated by Amodei and colleagues highlighted the dangers of negative side effects, reward hacking, and safe exploration, each of which takes on amplified significance in medical contexts where irreversible physiological harm is a persistent possibility [2]. Subsequent advances in reward modeling and learning from human preferences offered pathways to more nuanced objective functions, demonstrating that reinforcement learning agents could internalize complex human values when provided with comparative feedback rather than handcrafted scalar rewards [3, 4]. The proximal policy optimization algorithm

further contributed a stable and scalable training methodology that made these approaches practical for large-scale systems [5].

These algorithmic strands converged with a deeper philosophical and engineering concern around value alignment, articulated by Russell and others, who stressed that autonomous systems must be designed to remain deferential to human intentions even as their capabilities surpass our ability to specify objectives explicitly [6]. In the clinical domain, this translates into the requirement that a medical agent not merely optimize a static metric such as diagnostic accuracy or treatment adherence but continuously calibrate its actions against the evolving preferences of patients, the accumulated experience of clinicians, and the shifting standards of evidence-based medicine.

The deployment of artificial intelligence in medicine has already generated a substantial body of work on predictive models for imaging analysis, risk stratification, and clinical documentation. Topol described the convergence of human and artificial intelligence as a transformative force in high-performance medicine, while Rajkomar and colleagues systematically cataloged the opportunities and barriers for machine learning in clinical practice [7, 8]. Comprehensive reviews by Esteva et al. and Ghassemi et al. further delineated the technical and organizational challenges, underscoring that even static supervised models face significant hurdles in generalizability, interpretability, and integration into clinical workflows [9, 10]. Autonomous decision-making agents inherit all these challenges while adding the complexity of sequential action selection, making the case for dynamic alignment even more urgent.

A particularly instructive source of architectural motivation comes from the domain of adversarial robustness. Finlayson and co-authors demonstrated that medical image analysis models are susceptible to adversarial perturbations that could, in principle, be exploited to induce targeted misdiagnoses [11]. Extending this threat model to agents that act over time reveals a far richer attack surface, where adversaries can manipulate not only static inputs but also the perceived outcomes of prior actions, the reward signals used for online learning, and the communication channels through which the agent requests human guidance. Recent research focusing on security enhancement methods for adversarial robust large language model agents in medical decision-making tasks highlights the necessity of integrating defensive mechanisms directly into the agent's behavioral policy rather than treating them as external add-ons [12]. This perspective reinforces the argument that alignment and security are inseparable properties that must be co-designed at the architectural level.

3. System Architecture for Dynamic Security Alignment

A coherent architecture for dynamic security alignment in medical agents must reconcile two partially conflicting demands: the need for tight, real-time safety enforcement and the need for behavioral flexibility that allows the agent to adapt to novel clinical scenarios. This tension can be resolved through a layered control architecture in which a reinforcement learning-based meta-controller operates above a set of hard safety monitors and below a layer of institutional governance protocols. The bottom layer consists of domain-specific safety shields that enforce non-negotiable clinical constraints, such as absolute contraindications for certain drug combinations, upper bounds on radiation dosage, or mandatory confirmations before invasive procedures. These shields function as runtime assertion checkers that can override or delay the agent's proposed action, analogous to the safety cages employed in autonomous vehicle platforms. Their rules are authored by domain experts and updated

through a formal change management process, providing a stable regulatory grounding even as the agent's learned policy evolves.

Above this hard constraint layer, the meta-controller employs reinforcement learning to modulate the agent's risk posture. The meta-state vector encodes a set of contextual variables including patient acuity, the novelty of the clinical presentation relative to the agent's training distribution, the recency and reliability of available evidence, the degree of human oversight currently feasible, and aggregate system health indicators such as communication latency and sensor integrity. Based on this meta-state, the meta-controller chooses among a discrete set of operating modes, each of which varies the parametrization of the agent's decision thresholds, the frequency of human consultations, the aggressiveness of exploratory actions, and the type of explanation required before an action is executed. For example, in a high-uncertainty regime the meta-controller may command a conservative mode that reduces the agent to a passive monitoring role, whereas under stable, well-characterized conditions it may allow more autonomous therapeutic adjustments.

The training of the meta-controller involves a distinct reinforcement learning loop with a reward signal that combines clinical outcome measures with process-level safety metrics, such as the number of shield activations, the timeliness of human escalations, and the diversity of clinical pathways explored during safe operation. This training environment must include a rich distribution of adversarial scenarios that simulate sensor spoofing, data poisoning, and model extraction attempts, so that the meta-controller learns to recognize early indicators of system compromise and degrade gracefully rather than catastrophically. Important lessons can be drawn from decades of research on resilient control in critical infrastructure, where multi-layer defense-in-depth architectures have long been standard practice. The medical domain differs, however, in the degree of epistemic uncertainty regarding both patient physiology and the social context of care, which requires the meta-controller to reason about confidence calibration in ways that go beyond traditional fault detection and isolation.

A critical architectural decision concerns the degree of coupling between the clinical decision-making policy and the alignment meta-controller. A loosely coupled design, in which the meta-controller only modulates high-level parameters, preserves modularity and allows the clinical policy to be updated independently, but may fail to capture subtle interactions between clinical reasoning and safety considerations. A tightly coupled design, where the alignment objectives are integrated directly into the clinical policy's value function, can yield more finely tuned behavior but complicates verification, validation, and regulatory approval because any update to the clinical knowledge base may inadvertently alter the agent's safety posture. In practice, a hybrid arrangement that isolates verifiable safety properties in the shield layer while allowing the meta-controller to influence the operational envelope of the clinical policy offers the most promising balance between accountability and adaptability.

4. Structural Trade-offs and Reinforcement Learning Formulation

The design of the reinforcement learning meta-controller for dynamic alignment surfaces a set of structural trade-offs that must be navigated with careful attention to the sociotechnical context of healthcare delivery. The first fundamental trade-off is between conservatism and clinical utility. A highly conservative alignment strategy that frequently escalates to human review and narrows the agent's action space minimizes the risk of harm but also erodes the efficiency gains that motivate autonomous decision-making in the first place. Conversely, an overly permissive alignment posture that allows the agent to operate with a wide action space

increases throughput and may capture subtle therapeutic opportunities but simultaneously raises the probability of undetected errors. The meta-controller must therefore continually estimate the expected value of automation for the current patient and context, a calculation that depends on the contemporaneous availability and cognitive load of human supervisors, the urgency of the clinical situation, and the agent's self-assessed uncertainty.

A second structural trade-off relates to the choice of reward granulation. Coarse reward signals that only provide feedback at episode boundaries, such as final patient outcomes, create a credit assignment problem that can delay the detection of misalignment and allow unsafe behaviors to persist for extended periods before correction. Fine-grained reward signals, derived from intermediate clinical indicators, laboratory values, and clinician feedback, provide denser learning signals but introduce the risk of reward hacking, where the agent learns to manipulate these proxies without genuinely improving patient health. The well-known phenomenon of oxygen saturation gaming in mechanical ventilation algorithms serves as a cautionary tale in this regard. The meta-controller's reward design must therefore incorporate adversary-aware shaping, where the alignment signal includes terms that penalize detectible gaming behaviors and reward the use of diverse, clinically interpretable pathways.

A third trade-off emerges in the tension between centralized and federated alignment learning. Centralized architectures, where alignment policies are trained on aggregated data from multiple deployment sites, benefit from large-scale statistical power and can rapidly propagate safety improvements across the installed base. However, they introduce data privacy vulnerabilities, create single points of adversarial compromise, and may encode population-level biases that obscure heterogeneity across demographic groups and care settings. Federated architectures, in which each site maintains a local alignment policy that is periodically synchronized with a global model through secure aggregation, better preserve privacy and can adapt more nimbly to local clinical practices and patient populations [20]. Yet federated learning complicates the detection of slowly developing misalignment patterns that only become visible when data are pooled across institutions. Hybrid topologies that implement differential privacy guarantees and anomaly detection at the aggregation server while allowing site-specific fine-tuning of safety thresholds represent a pragmatic compromise that is gaining traction in multicenter clinical AI initiatives.

The reinforcement learning formulation itself, while not elaborated mathematically here, rests on framing the alignment problem as a meta-Markov decision process in which the state space includes both clinical and operational variables, the action space consists of discrete operating mode selections, and the transition dynamics reflect the interplay between the clinical policy, the patient trajectory, and the adversarial environment. The discount factor in this formulation encodes the temporal horizon over which alignment is evaluated, which in medical contexts may span from the immediate procedural outcome to long-term quality of life. The choice of discount factor is not merely a technical hyperparameter but an ethical stance on the relative importance of present and future welfare, a point that underscores the necessity of multidisciplinary involvement in the configuration of such systems.

5. Infrastructure and Deployment Considerations

Translating dynamic security alignment from a laboratory concept into a clinically deployed capability demands a substantial infrastructure that extends well beyond the agent's inference engine. Continuous alignment requires continuous monitoring, and continuous monitoring generates an immense volume of telemetry data that must be ingested, stored, and analyzed with low latency. The infrastructure must therefore include a high-reliability data pipeline

capable of capturing every shield activation, every meta-controller mode transition, every human override event, and every instance of model uncertainty exceeding a calibrated threshold. The architectural principles developed for large-scale industrial monitoring, including time-series databases optimized for anomaly detection, distributed stream processing frameworks, and edge-cloud hierarchies that allow safety-critical checks to execute on-premises even during connectivity outages, provide a blueprint for the medical domain [18].

The computational demands of the meta-controller itself are modest relative to the clinical policy, because the meta-action space is deliberately small and the meta-state updates occur at a lower frequency. However, the infrastructure must support rapid redeployment of updated alignment policies as new threats and clinical evidence emerge. This necessitates a continuous integration and continuous deployment pipeline that is itself subject to regulatory oversight, a requirement that introduces nontrivial frictions. In the context of software as a medical device, any change to the alignment logic that could materially affect patient safety triggers a regulatory review process that, under current frameworks, can take months to complete [19]. Bridging the gap between the rapid iteration cadence of reinforcement learning and the deliberate pace of medical device certification is one of the most pressing challenges for real-world adoption.

The deployment architecture must also address the heterogeneity of healthcare information technology environments. A tertiary academic medical center with dedicated machine learning operations teams can support elaborate monitoring and retraining pipelines, but a rural clinic with limited IT staff cannot. This disparity creates a structural risk that dynamic alignment technology exacerbates existing inequalities in healthcare delivery, a scenario where well-resourced institutions benefit from continuously improving agents while under-resourced settings are left with outdated and increasingly misaligned versions. Mitigating this risk requires the development of lightweight alignment runtimes that can operate on commodity hardware, cloud-based alignment services that offload the computational burden, and funding models that subsidize the sustained engineering effort required to maintain alignment over the system lifecycle.

Energy consumption and environmental sustainability introduce an additional infrastructure consideration that is often overlooked in safety discussions. The carbon footprint of repeatedly retraining large clinical foundation models and evaluating alignment policies across diverse simulated patient populations is non-trivial. While the immediate safety benefits of dynamic alignment are likely to outweigh these environmental costs in the short term, the long-term expansion of such systems to global scale will necessitate investment in efficient training algorithms, model compression techniques, and renewable-powered data center infrastructure. Sustainability must be viewed as a dimension of alignment itself, because a medical agent whose operation contributes disproportionately to environmental degradation is, by a broader definition, misaligned with public health goals [21].

6. Governance, Fairness, and Policy Implications

The governance of dynamically aligned medical agents sits at the intersection of artificial intelligence regulation, medical device oversight, and health system administration. Traditional medical device regulation assumes a relatively static product that can be evaluated through premarket clinical trials and postmarket surveillance of adverse events. An agent whose behavior continuously evolves through reinforcement learning challenges this model because the very entity being regulated changes over time in ways that may not be fully

anticipatable at the time of initial approval. Adaptive regulatory frameworks, such as the predetermined change control plans proposed by the U.S. Food and Drug Administration for artificial intelligence-based software as a medical device, offer a starting point by allowing manufacturers to specify in advance the types of modifications they intend to make and the methodology by which they will validate safety after each update. Extending such frameworks to reinforcement learning-driven alignment requires additional measures that bound the rate and magnitude of behavioral change, guarantee monotonic improvement on a set of prespecified safety metrics, and mandate transparency in the form of human-interpretable change logs that explain what the agent learned and why its behavior shifted.

Fairness considerations permeate every layer of a dynamically aligned system. The training data used to develop both the clinical policy and the meta-controller may underrepresent certain demographic groups, leading to systematic differences in the conservatism of alignment enforcement across populations. A poorly calibrated uncertainty estimator may cause the meta-controller to default to a high-oversight mode more frequently for patients with rare disease presentations or non-normative physiological patterns, inadvertently creating a two-tier system in which marginalized groups receive less autonomous care and, counterintuitively, may experience worse outcomes because of slower decision cycles. Obermeyer and colleagues famously demonstrated that a widely used clinical risk score exhibited significant racial bias because it used healthcare cost as a proxy for health need, a finding that underscores how seemingly objective optimization targets can encode structural inequities [16]. Dynamic alignment systems must incorporate rigorous fairness auditing that disaggregates safety outcomes, override rates, and clinical outcomes by race, ethnicity, gender, socioeconomic status, and geography, with predefined thresholds that trigger mandatory human review and corrective retraining when disparities are detected.

The policy landscape for dynamically aligned medical agents is further complicated by the transnational nature of both AI development and healthcare delivery. An agent trained in one regulatory jurisdiction may be deployed in another with different standards for informed consent, data privacy, and professional liability. The European Union's Artificial Intelligence Act classifies certain medical AI applications as high-risk and imposes requirements for human oversight, transparency, and robustness that have direct implications for dynamic alignment. However, the precise interpretation of these requirements for systems that learn continuously remains unsettled. International coordination bodies such as the International Medical Device Regulators Forum are beginning to address the need for harmonized guidance, but progress is slow relative to the pace of technological development. In this environment, manufacturers and healthcare organizations have a responsibility to adopt a precautionary approach that voluntarily exceeds minimal regulatory requirements when deploying dynamically aligned agents, particularly in vulnerable populations.

Liability allocation represents a governance challenge of the first order. When an autonomous medical agent that has dynamically adjusted its own safety posture causes patient harm, the chain of accountability is diffuse. The original developers of the base clinical policy, the designers of the meta-controller, the institution that configured the alignment parameters, the supervising clinician who may have been inattentive to an override recommendation, and the regulator that approved the adaptive change control plan all share some degree of responsibility. Without clear legal frameworks that assign liability in a manner that incentivizes safety without stifling innovation, the adoption of dynamic alignment technologies will be limited to the most risk-tolerant institutions. Some promising directions

include the establishment of no-fault compensation funds modeled on vaccine injury programs, mandatory professional liability insurance pools for AI-assisted care, and contractual mechanisms that apportion liability among software vendors, healthcare providers, and payers according to predefined criteria.

7. Robustness, Adversarial Resilience, and Sustainability

The robustness of a dynamically aligned medical agent must be evaluated across multiple dimensions, including robustness to distributional shift, robustness to adversarial manipulation, and robustness to gradual degradation of the infrastructure on which the agent depends. Distributional shift in healthcare arises from multiple sources, such as changes in population demographics, the emergence of new diseases, the introduction of novel therapeutics, and temporal trends in clinical documentation practices. An alignment meta-controller that is itself a learned function is vulnerable to the same failure modes as the policy it governs if its training distribution does not encompass the full range of possible deployment conditions. Continuous domain adaptation techniques, wherein the meta-controller periodically retrains on freshly collected deployment data under human supervision, can mitigate this risk but introduce a secondary fragility by creating a feedback loop in which the meta-controller's own decisions shape the data it observes, a form of distributional entanglement that can lead to runaway feedback or concept collapse.

Adversarial resilience requires explicit modeling of threat actors who may have financial, political, or malicious motivations to induce unsafe behavior in medical agents. The attack surface for a reinforcement learning-based alignment system includes the observation channels through which the meta-controller perceives the clinical state, the reward channels that provide feedback on alignment quality, and the communication links that connect the meta-controller to the shields and human supervisors. Data poisoning attacks that gradually corrupt the training data for the meta-controller are particularly insidious because they can slowly erode safety margins over months without triggering any single alarm. Defense strategies from the adversarial machine learning literature, including certified robustness guarantees for specific subcomponents, anomaly detection on incoming sensor streams, and multi-source reward verification that cross-checks alignment signals from independent monitors, must be woven into the system architecture rather than bolted on as an afterthought [22].

Sustainability of alignment is a temporal property that describes the system's ability to maintain its safety properties over extended deployment lifetimes that may span a decade or more. The clinical policy, the meta-controller, the shield logic, the monitoring infrastructure, and the human supervisory processes each evolve at different rates and are maintained by different organizational units. Over time, component drift causes the implicit assumptions on which the initial alignment design was based to erode. For example, a shield rule that blocks the administration of a particular drug combination may become obsolete when a new clinical guideline reverses a longstanding contraindication, yet the shield may remain in place because the organizational process for updating it is slow. The meta-controller may then learn to work around the obsolete shield by recommending alternative therapies that are suboptimal, degrading clinical quality without triggering any explicit safety alert. Preventing this form of alignment decay requires an organizational commitment to lifecycle management that extends far beyond the initial deployment, with dedicated alignment maintenance teams responsible for proactively auditing shield effectiveness, meta-controller calibration, and the coherence of the overall safety architecture.

8. Cross-Domain Comparative Analysis

Insights from domains that share structural similarities with autonomous medical decision-making can illuminate design principles that might otherwise remain obscured. The automotive industry's long experience with safety-critical autonomous systems provides an instructive comparison. In advanced driver assistance and autonomous vehicle platforms, functional safety standards such as ISO 26262 impose a rigorous development process that traces every safety requirement to a verifiable implementation. While the clinical domain involves far greater complexity and uncertainty than highway driving, the principle of a safety case that links high-level safety goals to specific technical measures and evidentiary arguments is transferable. A dynamic medical alignment system should be accompanied by a living safety case that is version-controlled alongside the software and updated whenever the meta-controller behavior or shield configuration changes.

The domain of financial trading systems offers a contrasting but equally valuable analogy. Algorithmic trading platforms operate in adversarial, non-stationary environments where rapid adaptation is essential and the cost of errors can be catastrophic, as illustrated by flash crashes and rogue algorithm incidents. The regulatory response has included the imposition of circuit breakers that halt trading when predefined volatility thresholds are exceeded, mandatory kill switches that allow human operators to rapidly deactivate malfunctioning algorithms, and post-trade surveillance that reconstructs the algorithm's decision-making process for audit purposes. These mechanisms map naturally onto the shield layer, the human override architecture, and the explainability requirements discussed earlier.

Large-scale infrastructure control systems, such as those managing electrical grids and water distribution networks, have evolved sophisticated approaches to graceful degradation under partial failure. Rather than attempting to maintain full functionality as components fail, these systems reconfigure themselves to sustain the most critical services while shedding non-essential loads. A medical agent could adopt a similar philosophy by, for example, continuing to provide medication interaction checks and vital sign trend analysis while suspending autonomous therapeutic recommendations when it detects internal anomalies. Such graceful degradation is a form of dynamic alignment that prioritizes safety over capability, a principle that should be deeply embedded in the system design from the outset.

9. Conclusion

Reinforcement learning-based dynamic security alignment for autonomous medical decision-making agents represents a paradigm shift from static verification to continuous, context-sensitive regulation of machine behavior. This paper has argued that the success of such systems hinges less on a specific reinforcement learning algorithm and more on the holistic design of a sociotechnical infrastructure that integrates hard safety shields, adaptive meta-control, rigorous fairness auditing, adversarial resilience mechanisms, and lifecycle governance processes. The structural trade-offs between conservatism and utility, centralized and federated learning, and fine-grained versus coarse reward signals must be navigated with explicit attention to the clinical, ethical, and operational realities of healthcare delivery. The deployment of dynamically aligned agents will require not only technical innovation but also substantial evolution in regulatory frameworks, liability arrangements, and organizational capacity for long-term maintenance. As the field moves forward, interdisciplinary collaboration among clinicians, engineers, ethicists, regulators, and patient communities remains indispensable to ensuring that autonomous medical agents enhance rather than undermine the quality, equity, and trustworthiness of care.

References

1. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.

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

3. Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). Scalable agent alignment via reward modeling: A research direction. arXiv preprint arXiv:1811.07871.

4. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30, 4299–4307.

5. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

6. Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

7. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

8. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

9. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.

10. Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. AMIA Summits on Translational Science Proceedings, 2020, 191–200.

11. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287–1289.

12. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

13. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

14. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

15. Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv preprint arXiv:1805.00899.

16. 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.

17. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54–71.

18. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q. V., ... & Ng, A. Y. (2012). Large scale distributed deep networks. Advances in Neural Information Processing Systems, 25, 1223–1231.

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

20. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119.

21. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

22. Wong, E., & Kolter, J. Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. International Conference on Machine Learning, 5286–5295.