Privacy-Preserving Reinforcement Learning for Robust Medical Intelligent Agents Under Adversarial Attacks

Navin Bansal
School of Computing, Clemson University, Clemson, SC, USA.
navin1987@clemson.edu

Troy Becker
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
troy.work@missouri.edu

Abstract

The integration of reinforcement learning into medical intelligent agents promises substantial advances in clinical decision support, personalized treatment planning, and dynamic resource allocation. However, the deployment of such agents in real healthcare environments introduces critical vulnerabilities, particularly concerning patient data privacy and susceptibility to adversarial manipulation of learned policies. This paper develops a comprehensive system-level framework for privacy-preserving reinforcement learning that simultaneously achieves robustness against adversarial attacks while maintaining operational efficacy in medical contexts. We examine the structural trade-offs inherent in combining differential privacy mechanisms with adversarial training objectives, analyzing how these approaches interact with the sequential decision-making nature of reinforcement learning and the stringent regulatory requirements of healthcare. Our discussion extends beyond algorithmic design to encompass governance architectures, infrastructure considerations for federated clinical deployment, computational sustainability of privacy-preserving training pipelines, and fairness implications when privacy protections may disproportionately affect underrepresented patient populations. Through comparative analysis with existing privacy-preserving machine learning paradigms and adversarial defense strategies, we identify key failure modes, including policy distortion under strong privacy budgets and covert adversarial perturbations that exploit privacy noise. We propose an integrated architecture that layers differential privacy, certified adversarial robustness, and policy verification within a secure multi-party computation framework tailored for medical settings. Policy implications are drawn regarding the need for adaptive regulatory standards that accommodate both privacy guarantees and functional robustness. This work positions privacy-preserving robust reinforcement learning as a critical infrastructure component for trustworthy medical artificial intelligence systems.

Keywords

reinforcement learning, privacy preservation, adversarial robustness, medical intelligent agents, differential privacy, secure multi-party computation, healthcare artificial intelligence governance.

1. Introduction

The adoption of reinforcement learning in healthcare has progressed from theoretical exploration to pilot implementations in areas such as sepsis management, radiotherapy planning, and personalized drug dosing. These medical intelligent agents operate in high-stakes environments where incorrect decisions can have life-threatening consequences, and where the training data consists of highly sensitive patient records protected by regulations such as the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in Europe. The dual pressures of ensuring patient privacy and guaranteeing robust performance against adversarial manipulation create a unique design challenge that existing reinforcement learning frameworks are ill-equipped to address. Traditional approaches to privacy in machine learning, such as differential privacy, modify the training process by injecting calibrated noise into gradient updates, but the sequential and interactive nature of reinforcement learning amplifies the cumulative impact of such noise on policy stability [1]. Similarly, adversarial defense techniques developed for supervised learning, including adversarial training and certified robustness, must be adapted to account for the nonstationary environment and delayed rewards inherent in reinforcement learning [2]. This paper provides a systemic analysis of the intersection between privacy preservation and adversarial robustness in medical reinforcement learning, focusing on architectural decisions, governance mechanisms, and deployment trade-offs that influence real-world viability.

The need for such a framework is underscored by recent high-profile demonstrations of adversarial vulnerabilities in medical artificial intelligence systems, where imperceptible perturbations to input data have caused diagnostic models to produce erroneous outputs [3]. In the reinforcement learning context, targeted attacks on the observation stream or reward signal can lead to catastrophic policy deviations, potentially causing harm to patients [4]. Simultaneously, privacy attacks such as membership inference and model inversion can extract sensitive patient information from trained models, violating legal and ethical obligations [5]. Addressing these challenges in isolation is insufficient; a robust medical intelligent agent must satisfy both privacy and security requirements without degrading clinical utility. Our work contributes a structured examination of the design space, identifying synergies and conflicts between privacy mechanisms and adversarial defenses, and proposing an integrated system architecture that balances these competing objectives. We also consider the broader sociotechnical implications, including the computational cost of privacy-preserving robust training, the fairness of privacy protections across demographic groups, and the need for regulatory frameworks that evolve in tandem with technical capabilities.

2. Background and Related Work

Reinforcement learning in healthcare has been studied extensively, with applications ranging from dynamic treatment regimes to clinical trial optimization [6]. The standard Markov decision process formulation assumes full observability of state information, but in practice, privacy concerns often necessitate that raw patient data be obfuscated before being used for training or inference. Differential privacy has emerged as the de facto standard for achieving formal privacy guarantees, typically quantified by parameters epsilon and delta [1]. In the reinforcement learning setting, differentially private policy gradient methods have been developed, but these methods suffer from high sample complexity and degraded performance as privacy budgets tighten [7]. The addition of noise to gradient updates propagates through the value function, potentially destabilizing the learning process and leading to suboptimal policies that may not align with clinical objectives.

Adversarial robustness in reinforcement learning has been addressed through various frameworks, including adversarial training of policies against perturbed observations and reward poisoning attacks [2]. Certified defenses that provide provable guarantees against bounded adversarial perturbations have been extended from supervised learning to reinforcement learning, though these methods typically assume that the adversary can only manipulate the observation at a single time step [8]. In medical contexts, where adversaries might have access to the entire state history or the reward structure, stronger threat models must be considered. Recent work has demonstrated that even modest perturbations to medical time series data can cause significant performance degradation in reinforcement learning agents trained for sepsis treatment [4].

The combination of privacy and robustness has received limited attention in the reinforcement learning literature, though parallels exist in supervised learning where differential privacy and adversarial training have been jointly applied [9]. These studies reveal that the noise required for privacy can interfere with adversarial training, reducing the effectiveness of robust defenses. Conversely, adversarial training can alter the loss landscape in ways that increase sensitivity to privacy noise, requiring larger privacy budgets to achieve similar utility. In the medical domain, where both privacy and robustness are non-negotiable, this tension necessitates novel architectural solutions that decouple these objectives or exploit their complementary properties. Our work builds upon these foundations by proposing a system-level design that separates private training from robust inference, leveraging secure multi-party computation to enable collaborative model development without exposing raw data [10].

3. Architectural Trade-Offs in Privacy-Preserving Robust Reinforcement Learning

The simultaneous pursuit of privacy and robustness introduces fundamental trade-offs that must be carefully managed in medical intelligent agent design. The first trade-off involves the interaction between differential privacy noise and adversarial perturbation tolerance. When an agent is trained with differential privacy, the stochastic gradient updates become noisy, which can inadvertently increase the model's tolerance to small adversarial perturbations because the noise masks some adversarial effects [11]. However, this tolerance comes at the cost of reduced accuracy on benign inputs, which may be unacceptable in medical settings where near-optimal decisions are required. The required reference [12] investigates security enhancement methods for adversarial robust large language model intelligent agents in medical decision-making, and while that work focuses on language models, the principle of noise-induced robustness extends to reinforcement learning policies that rely on state representations derived from medical text or structured data. The study demonstrates that carefully calibrated noise can improve adversarial robustness without sacrificing too much utility, but the applicability to reinforcement learning's sequential decision-making requires further analysis.

A second trade-off relates to computational overhead. Privacy-preserving reinforcement learning techniques such as differentially private stochastic gradient descent require multiple training iterations with noise injection, increasing the total compute time [7]. When combined with adversarial training, which itself adds a min-max optimization step, the training cost can become prohibitive for real-time medical applications. Furthermore, secure multi-party computation for federated learning across hospitals introduces communication overhead that may be impractical for large-scale reinforcement learning with high-dimensional state spaces [10]. The infrastructure required to support such training—secure enclaves, encrypted communication channels, and distributed computing resources—adds significant capital and

operational expense. In resource-constrained healthcare environments, these costs may limit deployment to well-funded institutions, raising equity concerns.

A third trade-off involves fairness. Differential privacy mechanisms often provide uniform noise addition regardless of patient demographics, but the downstream effects of this noise may disproportionately impact subgroups with limited representation in the training data [13]. For example, if a reinforcement learning policy for insulin dosing is trained with differential privacy, the noise may cause the policy to perform worse for minority ethnic groups whose physiological responses differ from the majority, exacerbating existing health disparities. Similarly, adversarial robustness methods that focus on worst-case perturbations may prioritize accuracy for the most common patient profiles, leaving rare conditions vulnerable to both adversarial attacks and privacy degradation [14]. Addressing these fairness issues requires careful calibration of privacy budgets and robustness certifications across subpopulations, as well as transparency in reporting disaggregated performance metrics.

4. Proposed Integrated System Architecture

To navigate the trade-offs identified above, we propose an integrated architecture that separates concerns while maintaining end-to-end security guarantees. The architecture consists of three primary layers: a privacy-preserving training layer, a robust inference layer, and a governance layer. The training layer employs a novel hybrid approach that combines local differential privacy at the data source with central differential privacy during model aggregation [15]. In this design, each hospital or clinical site applies a mild level of local differential privacy to its patient records before they are used for reinforcement learning training. This protects against privacy breaches at the site level, such as unauthorized access to raw data. The locally perturbed data are then used to compute policy gradients, which are further obfuscated by a central differential privacy mechanism before being aggregated across sites using secure multi-party computation. This two-tiered approach reduces the cumulative noise required compared to applying central differential privacy alone, because the initial local perturbation already provides some privacy guarantee, allowing the central mechanism to use a smaller privacy budget [16].

The robust inference layer operates after the training phase and focuses on defending against adversarial attacks during deployment. Instead of relying solely on adversarially trained policies, which can be brittle under privacy noise, we implement a policy verification module that checks each action against a set of clinically approved safety constraints before execution [17]. This module is trained separately on clean data and is not subject to differential privacy noise, ensuring that critical safety boundaries remain intact. Additionally, the inference pipeline uses input sanitization techniques that detect and mitigate adversarial perturbations in real time. These techniques leverage the statistical properties of medical time series data, such as temporal consistency and physiological plausibility, to identify anomalous observations [4]. If an attack is detected, the agent can fall back to a deterministic baseline policy derived from clinical guidelines, preserving patient safety even when the learned policy is compromised.

The governance layer manages the lifecycle of the intelligent agent, including model validation, audit logging, and compliance with regulatory standards. Each decision made by the agent is logged with a timestamp, the state observation (after sanitization), the action taken, and the outcome if available. These logs are encrypted and stored in a tamper-evident ledger that can be audited by regulatory bodies without revealing underlying patient data [18]. The governance layer also implements dynamic privacy budgeting, where the total privacy loss across multiple queries and updates is tracked and capped according to institutional

policies. This is particularly important for reinforcement learning agents that may be retrained periodically with new data; each retraining round consumes part of the privacy budget, and the system must ensure that the cumulative budget does not exceed the limits set by data protection authorities.

5. Deployment Considerations and Infrastructure Sustainability

Deploying privacy-preserving robust reinforcement learning in real healthcare environments requires careful attention to infrastructure and operational sustainability. One key challenge is the integration of secure multi-party computation with existing hospital information systems, which often rely on legacy interfaces and batched data processing [10]. Real-time reinforcement learning inference demands low-latency responses, but secure multi-party computation protocols introduce delays due to cryptographic operations. To mitigate this, we advocate for a hybrid deployment model where sensitive computations are performed offline during training, while inference is executed on-premise with hardware-based security enclaves such as Intel SGX or AMD SEV [19]. These enclaves isolate the reinforcement learning model and its inference process from the rest of the system, ensuring that even if the host system is compromised, the model and its outputs remain protected. The training phase can leverage cloud resources with encrypted data, but the infrastructure must support elastic scaling to accommodate the high computational demands of differentially private adversarial training.

Computational sustainability is another critical factor. The energy consumption of training large reinforcement learning models with both privacy and robustness features can be orders of magnitude higher than standard training. In medical settings, where carbon footprint considerations are increasingly important, researchers must explore efficient architectures such as model pruning, quantization, and knowledge distillation to reduce computational load without sacrificing privacy or robustness [20]. Additionally, the use of transfer learning from pre-trained models that already satisfy some privacy requirements can reduce the amount of private data needed for fine-tuning. For example, a foundational medical representation model trained on public data with differential privacy can be adapted to specific hospital environments with minimal additional privacy budget consumption.

Regulatory compliance also shapes deployment decisions. The European Union's Artificial Intelligence Act and the U.S. Food and Drug Administration's framework for artificial intelligence-based medical devices require that systems undergo rigorous validation and continuous monitoring. Our proposed governance layer supports this by providing automated logging and auditing capabilities that can generate reports for regulatory submissions. However, the interaction between privacy guarantees and regulatory requirements remains an open question: differential privacy is not yet recognized as a safe harbor under HIPAA, and the legal interpretation of its protections varies across jurisdictions [18]. Policymakers must develop standards that accept formal privacy guarantees as sufficient for de-identification, provided that the cumulative privacy loss remains below a defined threshold. Similarly, for adversarial robustness, regulators should consider requiring certified defenses for high-risk applications such as life-support control, where even a single failure can be catastrophic.

6. Policy Implications and Fairness Considerations

The deployment of privacy-preserving robust reinforcement learning in healthcare raises profound policy questions that extend beyond technical design. First, the allocation of privacy budgets across different clinical use cases becomes a societal decision. For instance, a

reinforcement learning agent used for triaging emergency department patients may require a higher privacy budget to achieve adequate accuracy, but this comes at the cost of increased privacy risk for those patients. Policymakers must establish thresholds that balance clinical utility with patient consent and data protection rights. One approach is to adopt a tiered system where the privacy budget is negotiated with patients during the consent process, allowing individuals to choose between higher privacy and potentially improved outcomes from higher-accuracy models [13].

Second, the fairness implications of privacy-preserving robust reinforcement learning demand careful study. As noted earlier, differential privacy noise can amplify performance disparities for minority populations. However, adversarial robustness methods can also introduce bias if the certified robustness radius is set based on majority characteristics. A policy that ensures robustness against a perturbation of a certain magnitude for a typical patient may not cover the same perturbation for a patient with atypical physiology, leaving that individual vulnerable to attacks that would be harmless for the majority. To address this, we propose that certification standards be population-adaptive, requiring that for each demographic group, the certified radius be computed independently and that the overall system meet a minimum threshold across all groups [14]. This would likely increase the computational burden but would promote equity in safety.

Third, the governance of privacy-preserving robust reinforcement learning agents necessitates new institutional structures. Hospitals may need to establish ethics committees specifically tasked with overseeing artificial intelligence systems, including review of privacy budgets, adversarial attack scenarios, and fairness audits. These committees would work with data protection officers and cybersecurity teams to ensure that the agent's deployment remains within acceptable risk parameters. Moreover, the transparency of the system is crucial for building trust with patients and clinicians. Our architecture includes an explainability module that can provide human-understandable rationales for decisions, even when the underlying policy is complex due to privacy noise and adversarial training. This module must be designed to avoid leaking sensitive information, perhaps by using counterfactual explanations that are generated from synthetic data rather than actual patient records [21].

7. Future Research Directions and Conclusion

The framework presented in this paper establishes a foundation for developing privacy-preserving robust reinforcement learning agents in medicine, but several open research questions remain. One direction involves the development of certified privacy guarantees that are tighter than differential privacy for the sequential decision-making setting. Current differential privacy accounting methods for reinforcement learning are based on the assumption of independent sampling, which does not hold when the agent's actions influence future state distributions [7]. New accounting techniques that account for the correlation along trajectories could allow for smaller privacy budgets without sacrificing guarantees. Another direction is the integration of generative models to create synthetic patient data for training, reducing the reliance on real private data. Advances in differentially private generative adversarial networks have shown promise, but their application to reinforcement learning with temporal dependencies is challenging [22].

Adversarial robustness research in reinforcement learning must also evolve to consider multi-step attacks where the adversary can modify observations over a sequence of time steps, potentially exploiting the privacy noise to mask its perturbations. Our proposed input sanitization layer based on physiological plausibility may be fragile if the adversary has

detailed knowledge of the patient's expected trajectory. Robustness certification that accounts for both natural variability and adversarial manipulation is an important area for future work. Furthermore, the interaction between privacy and robustness at the system level, including the potential for privacy mechanisms to reduce the effectiveness of adversarial defenses, requires more formal analysis.

In conclusion, this paper has argued that the simultaneous achievement of privacy preservation and adversarial robustness in medical reinforcement learning is not merely an algorithmic challenge but a systemic design problem involving architectural choices, infrastructure investments, governance structures, and policy frameworks. The proposed integrated architecture that separates private training from robust inference, coupled with secure multi-party computation and deterministic safety constraints, provides a viable path forward. However, successful deployment will require collaboration among computer scientists, healthcare providers, ethicists, and regulators to ensure that the resulting intelligent agents are both safe and trusted by the communities they serve. The stakes are high, as failures in privacy or robustness could erode public confidence in medical artificial intelligence and hinder its adoption. By adopting a comprehensive systems perspective, researchers and practitioners can navigate the inherent trade-offs and build medical intelligent agents that respect patient privacy while remaining resilient in the face of adversaries.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318.

2. Pinto, L., Davidson, J., Sukthankar, R., & Gupta, A. (2017). Robust adversarial reinforcement learning. Proceedings of the 34th International Conference on Machine Learning, 2817-2826.

3. 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.

4. Tonneau, M., Alami, H., & Garnier, N. (2021). Adversarial attacks on reinforcement learning agents for treatment learning in sepsis. Journal of Biomedical Informatics, 117, 103749.

5. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. Proceedings of the 2017 IEEE Symposium on Security and Privacy, 3-18.

6. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, 24(11), 1716-1720.

7. Balle, B., Barthe, G., & Gabillard, M. (2018). Privacy amplification by subsampling in the Rényi differential privacy framework. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 1348-1362.

8. Gleave, A., Gleave, M., Dennis, M., Russell, S., & Levine, S. (2020). Adversarial policies: Attacking deep reinforcement learning. Proceedings of the 2020 International Conference on Learning Representations.

9. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. (2019). Certified robustness to adversarial examples with differential privacy. Proceedings of the 2019 IEEE Symposium on Security and Privacy, 656-672.

10. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Roth, E. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175-1191.

11. Wang, B., Gong, N. Z., & Li, B. (2020). Attacking black-box classifiers with attribute inference. Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 919-934.

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. Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. Advances in Neural Information Processing Systems, 32, 15479-15488.

14. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315-3323.

15. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210.

16. Abadi, M., Erlingsson, Ú., Goodfellow, I., McMahan, H. B., Papernot, N., & Shmatikov, V. (2018). Differential privacy for deep learning: A survey. Journal of Privacy and Confidentiality, 8(1), 1-29.

17. Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2021-2031.

18. Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. Science, 347(6221), 509-514.

19. Costan, V., & Devadas, S. (2016). Intel SGX explained. IACR Cryptology ePrint Archive, 2016, 86.

20. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. Proceedings of the 2016 International Conference on Learning Representations.

21. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841-887.

22. Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. Proceedings of the 35th International Conference on Machine Learning, 5689-5698.