Position Paper: Rethinking Privacy in RL for Sequential Decision-making in the Age of LLMs

Flint Xiaofeng Fan
Centre for Frontier AI Research (CFAR)
Institute of High Performance Computing (IHPC)
Agency for Science, Technology and Research
Singapore
fxf@u.nus.edu

Cheston Tan
Centre for Frontier AI Research (CFAR)
Institute of High Performance Computing (IHPC)
Agency for Science, Technology and Research
Singapore
cheston-tan@i2r.a-star.edu.sg

Roger Wattenhofer
D-ITET
ETH Zurich
Zurich, Switzerland
wattenhofer@ethz.ch

Yew-Soon Ong
CCDS, Nanyang Technological University
CFAR, IHPC, Agency for Science, Technology and Research
Singapore
asysong@ntu.edu.sg

***Abstract*—The rise of reinforcement learning (RL) in critical real-world applications demands a fundamental rethinking of privacy in AI systems. Traditional privacy frameworks, designed to protect isolated data points, fall short for sequential decision-making systems where sensitive information emerges from temporal patterns, behavioral strategies, and collaborative dynamics. Modern RL paradigms, such as federated RL (FedRL) and RL with human feedback (RLHF) in large language models (LLMs), exacerbate these challenges by introducing complex, interactive, and context-dependent learning environments that traditional methods do not address. In this position paper, we argue for a new privacy paradigm built on four core principles: multi-scale protection, behavioral pattern protection, collaborative privacy preservation, and context-aware adaptation. These principles expose inherent tensions between privacy, utility, and interpretability that must be navigated as RL systems become more pervasive in high-stakes domains like healthcare, autonomous vehicles, and decision support systems powered by LLMs. To tackle these challenges, we call for the development of new theoretical frameworks, practical mechanisms, and rigorous evaluation methodologies that collectively enable effective privacy protection in sequential decision-making systems.**

***Index Terms*—Privacy, Reinforcement Learning, Sequential Decision-making, RLHF, LLMs**

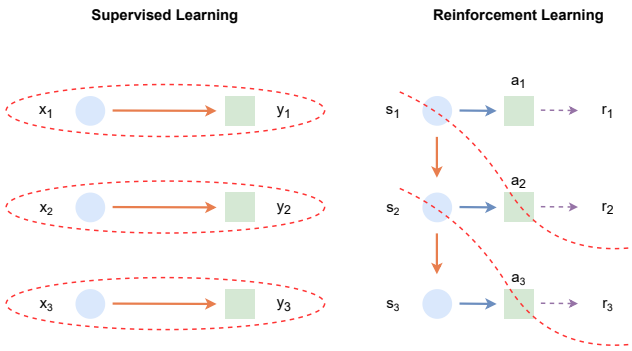

Fig. 1. **The essential contrast in privacy requirements between supervised learning (left) and reinforcement learning (right)**. In supervised learning, data points (x, y) are isolated and can be protected individually through local privacy mechanisms (red dashed ellipses). In reinforcement learning, a sequence of states s connected by actions a and rewards r creates temporal dependencies, highlighted by the red dashed curve that spans multiple decision points. While local privacy protection can be adapted in reinforcement learning, the sequential nature of state transitions and action-reward pairs creates dependencies that make point-wise privacy protection insufficient.

I. Introduction

The rise of reinforcement learning (RL) in critical real-world applications [1]–[5] has exposed a fundamental tension in AI privacy: How do we protect sensitive information in systems that learn and make decisions over time? Traditional privacy frameworks, built for protecting individual data points in static datasets [6]–[8], are increasingly inadequate for modern RL systems where sensitive information exists not just in individual moments but in temporal patterns, behavioral strategies, and collaborative dynamics [9]. These privacy challenges arise directly from RL's fundamental characteristic of learning through sequential interaction.

Unlike traditional machine learning paradigms, reinforcement learning operates through continuous interaction between an agent and its environment [2]. The RL agent observes the current state of the environment, takes actions based on these observations, and receives feedback in the form of rewards. This sequential learning process fundamentally differs from supervised learning, where most existing privacy frameworks were developed [6], [7]. As illustrated in Fig. 1, supervised learning data points are typically treated as independent samples, allowing privacy mechanisms to protect each point individually. However, RL violates this independence assumption, giving rise to three fundamental privacy concerns: First, *temporal patterns* emerge from sequential relationships between states, actions, and rewards, creating dependencies that span entire trajectories and potentially revealing sensitive

information about the underlying process [10]. Second, *behavioral strategies* develop as the agent learns optimal policies, encoding complete decision-making patterns that go beyond the simple input-output mappings of supervised learning and may reveal proprietary algorithms or institutional expertise [11]. Third, *collaborative dynamics* arise from the continuous adaptation between the agent and its environment, creating ongoing relationships that have no parallel in traditional supervised learning and potentially exposing sensitive information through patterns of response and adjustment [12].

While these privacy challenges are fundamental to basic RL systems, the emergence of advanced paradigms has further amplified these concerns, particularly in relation to collaborative dynamics. Federated reinforcement learning (FedRL), where multiple agents share learning experiences while keeping data locally [13]–[20], introduces the challenge of protecting not only individual agent data but also emergent collective behavioral patterns. Similarly, large language models (LLMs), such as ChatGPT [21] and DeepSeek [22], refined through reinforcement learning with human feedback (RLHF) [23]–[25] extend these collaborative privacy challenges to human-AI interaction, creating additional vulnerabilities around protecting annotator characteristics and cultural information encoded in feedback patterns [26], [27]. Recent analysis has uncovered numerous instances of personal data in publicly available RLHF datasets that had evaded removal [28], highlighting how even carefully curated training data can expose private user information. These advanced paradigms demonstrate how the temporal and behavioral aspects of privacy intertwine with collaborative dynamics, pushing privacy challenges beyond individual agent privacy to encompass group-level patterns and societal concerns.

Recent privacy regulations like GDPR [29] and HIPAA [30] establish strict requirements for protecting such sensitive information, but their frameworks—designed primarily for static data protection—struggle to address these dynamic aspects of RL systems. These regulations presume a clear distinction between protected and non-protected data, a distinction that blurs in RL systems where sensitive information often emerges from patterns of interaction rather than residing in individual data points. This fundamental mismatch between regulatory frameworks and the nature of RL systems creates significant challenges for deployment in regulated domains.

To address these challenges, this position paper:

1) Articulates why traditional privacy frameworks fundamentally fail for sequential decision-making systems
2) Proposes four core principles for a new privacy paradigm: multi-scale protection, behavioral pattern protection, collaborative privacy preservation, and context-aware adaptation
3) Identifies critical open problems and research directions for realizing effective privacy in sequential settings

The rest of this paper is organized as follows: Section II reviews the evolution of privacy approaches in sequential settings. Section III explains why traditional frameworks fail. Section IV proposes four core principles that leads to our *Sequential Privacy* framework for rethinking privacy in RL setting. Section V examines implications through real-world applications. Section VI outlines research directions, and we conclude in Section VII with a call for community action toward developing privacy frameworks that can meet the needs of modern RL systems.

II. Evolution of Privacy Approaches in Sequential Settings

Before examining why traditional approaches fail, we trace the historical development of privacy mechanisms and their attempts to address sequential decision-making contexts. This evolution reveals a progression of increasingly sophisticated approaches, each trying to overcome the limitations of its predecessors while inadvertently highlighting deeper challenges.

A. Technical Foundations

The field of privacy-preserving machine learning began with differential privacy (DP), introduced by Dwork et al. [6]. This framework provides a mathematical foundation for quantifying information leakage: a randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if changes to individual data points have only limited impact on the output distribution. Formally:

For any two *adjacent* datasets D and D' (differing in at most one data record), and for all measurable subsets S of possible outputs, $\mathcal{M}$ satisfies (ϵ, δ)-DP if:

$$\Pr[\mathcal{M}(D) \in S] \;\leq\; e^{\epsilon} \Pr[\mathcal{M}(D') \in S] \;+\; \delta.$$

Here, ϵ (the "privacy budget") controls the multiplicative gap in probabilities, while δ bounds the probability of a larger deviation. Smaller ϵ and δ imply stronger privacy guarantees.

Two properties made this framework particularly attractive:

- *Composition*: Privacy guarantees combine predictably over multiple analyses of the same dataset.
- *Post-processing*: Privacy guarantees persist under any data-independent transformation of the mechanism's output.

These properties proved highly effective for static data but would later reveal fundamental limitations in sequential settings.

B. Early Adaptations to Sequential Data

The first attempts to apply privacy to sequential settings emerged in the early 2010s, as researchers began working with temporal data like reinforcement learning trajectories:

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_T)$$

Early approaches split into two camps: those applying noise independently at each timestep, and those treating entire trajectories as atomic units. Both approaches revealed concerning trade-offs between privacy guarantees and system utility. Building on the original DP framework, later work extended the ideas to streaming and continual observation settings [31], [32] and further explored private sequential learning in online contexts [33] and with bandit feedback [34]. These efforts highlighted a fundamental tension: the very temporal

correlations that make sequential learning effective also create new privacy vulnerabilities.

C. The Cryptographic Era

As limitations of noise-based approaches became apparent, the field of privacy-preserving learning shifted toward cryptographic solutions. Researchers explored secure multi-party computation and homomorphic encryption, aiming to enable secure computation without data sharing [35]. At the same time, privacy-preserving deep learning emerged, with early frameworks showing how collaborative deep models could be trained without directly sharing sensitive data [36], [37]. While these approaches provided strong cryptographic guarantees, they faced significant scalability challenges and, more fundamentally, couldn't address the broader issue of behavioral pattern privacy that emerges in sequential settings [12].

D. Information-Theoretic Approaches

The late 2010s saw researchers turn to information theory, introducing mutual information constraints to limit information leakage through learned policies [38], [39]. This marked an important shift in thinking—from protecting individual data points to considering the information content of behavioral patterns. While these approaches provided new theoretical insights, they highlighted the difficulty of balancing privacy with the need to preserve useful temporal patterns.

E. Modern Developments

Recent years have seen two parallel developments that further complicate the privacy landscape. On one front, advances in deep learning under differential privacy have been refined and deployed at scale [7], [40]. These works leverage advanced privacy accounting (e.g., the moment accountant) to tightly track cumulative privacy loss during iterative training, ensuring high utility despite strict privacy constraints. On another front, theoretical insights such as privacy amplification by iteration have demonstrated that the effective privacy loss in iterative algorithms can be significantly reduced [41]. Moreover, modern private learning approaches continue to grapple with the challenges of gradient inversion and recovering sensitive information from model updates [39], [42].

III. Why Traditional Privacy Approaches Fail

The historical evolution of privacy approaches reveals not just technical limitations but fundamental incompatibilities with sequential decision-making. Here we analyze why these approaches fail, showing that the challenges arise from core properties of sequential learning rather than implementation limitations.

A. The Sequential Nature of RL

Consider the structure of RL data: a continuous stream of interactions between an agent and its environment, generating trajectories of the form

$$\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_T)$$

where each state-action pair influences the entire future sequence. As shown in Fig. 1, unlike supervised learning data, which is often (implicitly) assumed to be independent and identically distributed, RL trajectories exhibit strong temporal dependencies [43].

Because decisions at each timestep impact future states and rewards, RL reveals sensitive information not simply at isolated moments but in multi-step patterns and dynamic behaviors. As a result, privacy threats can emerge in unforeseen ways. For example, small changes in an action sequence can cascade and expose strategic aspects of a policy, or aggregated trajectories across agents can leak private information about a collective population. These challenges do not arise merely from implementation details; rather, they stem from the fundamental *sequential* nature of RL.

B. Exploration-Exploitation Privacy Vulnerabilities

The exploration-exploitation dilemma—central to RL but absent in supervised learning—introduces additional privacy concerns [47]. During exploration, an agent must try different actions to discover optimal strategies, creating behavioral patterns that can leak sensitive information about:

- *Knowledge boundaries*: Exploration patterns reveal what an agent doesn't know, potentially exposing gaps in training data
- *Learning dynamics*: The transition from exploration to exploitation creates temporal signatures that reflect training procedures
- *Uncertainty profiles*: Methods using uncertainty-based exploration (e.g., UCB algorithms) directly expose confidence estimates derived from private data

As a consequence, traditional privacy mechanisms face a new challenge: adding sufficient noise to mask exploration patterns can severely impair learning efficiency, while preserving learning efficiency may reveal sensitive information through exploration behavior. This tension exacerbates the already complex privacy-utility tradeoff in RL.

C. Analysis of Core Limitations

Building on the above observations, we identify four core limitations that make traditional privacy approaches fundamentally inadequate for RL systems.

1) Temporal Privacy Challenge: The *Temporal Privacy Challenge* arises from the fact that each timestep in an RL trajectory is tightly coupled with past and future timesteps. As shown by Mironov [48] and Zhang et al. [44], privacy loss can grow faster than standard composition theorems would predict, because small inferences at one timestep accumulate into bigger insights about the entire trajectory. In other words, even if each individual (s_t, a_t) pair is "protected" in isolation, long-range temporal correlations can still reveal private information. This challenge implies that privacy cannot be guaranteed by simply masking individual points; we must consider correlations across multiple temporal scales.

TABLE I
SUMMARY OF CORE PRIVACY CHALLENGES IN SEQUENTIAL DECISION-MAKING

Privacy Challenge	Root Cause	Prior Work Attempts	Limitations
Temporal	Multi-step correlations across trajectory segments	[31], [32], [44]	Standard composition theorems insufficient; privacy loss accumulates faster than expected
Behavioral	Learned policy encodes sensitive information in behavioral patterns	[11], [9]	Per-sample protection fails to bound policy-level leakage; dynamic sensitivity undermines guarantees
Collaborative	Non-local information flow across agents and human feedback	[12], [42], [27]	Local protection mechanisms ignore global patterns; aggregated updates expose institutional strategies
Context-Dependent	Domain-specific constraints and varying regulatory requirements	[45], [29], [30], [46]	One-size-fits-all frameworks inadequate for diverse deployment contexts

2) Behavioral Privacy Challenge: The *Behavioral Privacy Challenge* stems from *dynamic sensitivity* and the possibility that a learned policy encodes sensitive information in its behavioral patterns [9], [49]. A shift in a single action can propagate through future states and unravel details about the underlying policy (or the environment that shaped that policy). Cundy et al. [11] show how observing an agent's behavior distribution can leak crucial details about training data or proprietary algorithms. Traditional privacy frameworks that focus on per-sample data protection cannot bound this "policy-level" leakage. In domains where the strategy itself is sensitive (e.g., proprietary approaches in healthcare or competitive settings), such behavioral leakage is a direct threat.

3) Collaborative Privacy Challenge: The *Collaborative Privacy Challenge* emerges from multi-agent systems, federated reinforcement learning (FedRL), and human-in-the-loop methods (RLHF). These settings involve continuous sharing of updates, observations, or feedback among different parties (agents, servers, humans). Even if each local dataset or feedback instance is protected, the *aggregated patterns* of interaction and adaptation can reveal sensitive information about individuals or institutions [12], [42]. This non-local information flow is especially difficult to secure, because high-level behaviors or group updates can expose private attributes that traditional pointwise protection ignores.

4) Context-Dependent Privacy Challenge: Finally, the *Context-Dependent Privacy Challenge* arises from the fact that privacy sensitivities and regulatory requirements vary drastically across different domains, user populations, and deployment contexts. A healthcare RL system must comply with HIPAA or GDPR, while an autonomous vehicle system may face different proprietary or safety-driven constraints [45]. Similarly, an LLM refined via RLHF may need to guard the cultural or demographic information of human annotators in ways that differ from other RL use cases. Privacy is not "one size fits all," and the severity of temporal, behavioral, and collaborative leakage depends intimately on the context in which the RL system is deployed. This challenge calls for *adaptive* privacy frameworks that respond to domain- or population-specific requirements.

D. Failed Extension Attempts

Attempts to extend traditional privacy frameworks to RL reveal fundamental limitations. Adding noise at individual timesteps often destroys sequential structure, while protecting entire trajectories sacrifices utility. Cryptographic approaches [35] provide strong guarantees but face prohibitive computational overhead and still struggle with higher-level behavioral patterns.

Sophisticated techniques in federated or collaborative settings cannot fully prevent leakage of global patterns, as demonstrated by gradient inversion attacks [42] and adaptive interactions revealing participant attributes [12]. Recent work [50], [51] shows progress under restricted settings but still fails to address the interplay of temporal, behavioral, collaborative, and contextual factors.

E. Implications for Privacy Design

Taken together, these observations emphasize that privacy in RL cannot be retrofitted with simple modifications of traditional frameworks. The four challenges require us to:

1) *Address multi-step correlations* (temporal challenge) $\rightarrow$ **Multi-scale Privacy Protection** (Section IV-A)
2) *Protect the policy itself* (behavioral challenge) $\rightarrow$ **Behavioral Pattern Protection** (Section IV-B)
3) *Preserve privacy across interacting agents and humans* (collaborative challenge) $\rightarrow$ **Collaborative Privacy Preservation** (Section IV-C)
4) *Adapt protections to domain-specific constraints* (context-dependent challenge) $\rightarrow$ **Context-Aware Adaptation** (Section IV-D)

Any new approach must *jointly* tackle these dimensions while striking a careful balance between privacy, utility, interpretability, and feasibility. These requirements motivate the principles we propose next.

IV. CORE PRINCIPLES FOR SEQUENTIAL PRIVACY

We now introduce four core principles that address the challenges identified in Section III.

A. Multi-scale Privacy Protection

To counteract the *Temporal Privacy Challenge*, privacy must hold across multiple temporal scales—not just at the granularity of individual actions or states. We extend traditional definitions of differential privacy to account explicitly for trajectory segments of varying lengths [44], [48]:

Definition 4.1 (Multi-scale Privacy): A mechanism $\mathcal{M}$ provides (k, ϵ, δ)-multi-scale privacy if for all scales $1 \leq j \leq k$ and all trajectory segments $\tau_{t:t+j}$,

$$\Pr[\mathcal{M}(\tau_{t:t+j}) \in S] \;\leq\; e^{\epsilon_j}\, \Pr[\mathcal{M}(\tau'_{t:t+j}) \in S] \;+\; \delta_j,$$

where $\tau_{t:t+j}, \tau'_{t:t+j}$ are adjacent trajectory segments of length j, and ϵ_j (respectively δ_j) may increase with segment length.

Such multi-scale protection ensures that RL trajectories do not leak information cumulatively over time, thereby limiting the adversary's ability to reconstruct sensitive patterns from sequential data. It generalizes the usual composition theorems in differential privacy [6] by allowing an ϵ_j budget for each trajectory segment length j. In practice, one might set ϵ_j to grow sub-linearly in j to reflect partial reuse of noise across overlapping segments, or adopt advanced composition results that limit how quickly the overall privacy budget depletes over time. A rigorous analysis requires bounding correlations between overlapping segments $\tau_{t:t+j}$, which is an open theoretical question. For instance, one could assume a Markov property and then derive ϵ_j by combining concentration inequalities with standard DP composition results–however, the exact rate of growth in ϵ_j would depend on the mixing time of the underlying Markov chain. Investigating these parameter choices remains an important research direction.

B. Behavioral Pattern Protection

Addressing the *Behavioral Privacy Challenge* requires protecting the *policy*—i.e., the mapping from states to actions—rather than just the individual samples. This protection must cover both *exploitation patterns* (revealing what was learned) and *exploration patterns* (revealing uncertainty and learning dynamics). We thus focus on bounding divergences between entire trajectory distributions induced by different policies [11]:

Definition 4.2 (Behavioral Pattern Privacy): A policy learning mechanism $\mathcal{M}$ satisfies (α, β)-behavioral privacy if for any policies π_1, π_2 learned from adjacent training sets,

$$D_\alpha\big(\mathbb{P}_{\tau\sim\pi_1} \,\|\, \mathbb{P}_{\tau\sim\pi_2}\big) \;\leq\; \beta,$$

where D_α is the Rényi divergence [48] of order α and $\mathbb{P}_{\tau\sim\pi}$ is the distribution of trajectories under policy π.

This definition encompasses both exploitation and exploration behaviors, as π contains both components. However, exploration patterns present unique privacy challenges: they directly reveal uncertainty estimates which are typically derived from training data distributions. For example, in algorithms using upper confidence bounds (UCB) or Thompson sampling, the exploration strategy directly exposes confidence intervals calculated from private data.

By bounding how much a single agent's (or institution's) policy distribution—including both exploitation and exploration components—can shift under small changes in the underlying data, we reduce the risk that adversaries infer proprietary strategies, specialized treatment protocols, data sparsity patterns, or other policy-level knowledge.

C. Collaborative Privacy Preservation

The *Collaborative Privacy Challenge* is especially evident in federated or multi-agent RL, and in RLHF where human feedback is continuously integrated. We can frame this via information-theoretic constraints on collaborative systems [38]:

Definition 4.3 (Collaborative Privacy): A mechanism $\mathcal{M}$ provides (γ, η)-collaborative privacy if for all interaction histories H_t and new interactions i_t:

$$I(\mathcal{M}(H_t \cup \{i_t\}); \text{sensitive}_t | H_t) \leq \gamma$$

with probability at least $1 - \eta$, where $I(\cdot;\cdot|\cdot)$ denotes conditional mutual information and sensitive_t represents any sensitive attributes at time t including:

- Demographic information of participants
- Group-level behavioral patterns
- Institutional strategies or protocols
- Collective learning dynamics

This definition limits how much *additional* information is revealed about sensitive attributes (e.g., user demographics, group-level strategies) from each incremental interaction, even when prior interactions are already known.

D. Context-Aware Adaptation

Finally, the *Context-Dependent Privacy Challenge* demands that privacy guarantees adapt to different domains, user populations, and regulatory environments. We capture this adaptivity via:

Definition 4.4 (Context-Aware Privacy): A privacy mechanism $\mathcal{M}$ is (Θ, λ)-context-aware if for all contexts $c \in \mathcal{C}$ and privacy requirements $\theta_c \in \Theta$:

$$\Pr\left[\mathcal{M}(\tau, c) \text{ satisfies } \theta_c\right] \;\geq\; 1 - \lambda,$$

where θ_c specifies context-specific privacy parameters.

In high-stakes environments (e.g., clinical healthcare), θ_c may demand stricter bounds and narrower noise budgets, while less-sensitive tasks can tolerate relaxed protections. The mechanism adjusts its privacy parameters or noise injection strategies according to these evolving contextual requirements. It formalizes the idea that privacy mechanisms must adapt to different contexts while maintaining guaranteed levels of protection [8].

For instance, in a hospital environment, context-aware privacy might mean enforcing tighter privacy budgets (ϵ) for particularly sensitive patient attributes, in compliance with HIPAA or GDPR, while still allowing less-protected telemetry data to facilitate real-time decision-making. By contrast, in autonomous vehicles, the context might revolve around location data and proprietary driving logs: the system could relax certain bounds for purely operational metrics (e.g., mechanical sensors), but apply stricter protections for route data or user identities. These domain-specific variations underscore why static, one-size-fits-all privacy mechanisms often fail in practice: each context demands unique trade-offs between privacy, regulatory compliance, and system performance.

E. The Sequential Privacy Framework

These four principles form our *Sequential Privacy Framework* for reinforcement learning in sequential decision-making. A straightforward way to ensure that all aspects of privacy are satisfied is to enforce each principle independently and then take a worst-case (intersection) view of the guarantees. Formally, the overall privacy guarantee can be expressed as:

$$\mathcal{P}(\tau) = \min_{i \in \{1,2,3,4\}} \mathcal{P}_i(\tau),$$

where $\mathcal{P}_i(\tau)$ represents the privacy guarantee derived from each of the four principles (multi-scale, behavioral, collaborative, and context-aware). This worst-case perspective provides a conservative baseline, ensuring that an adversary cannot exploit any single dimension of leakage.

In practice, implementing these principles may involve combining multiple mechanisms:

$$\mathcal{M}(\tau) = h\big(\mathcal{M}_{\text{multi}}(\tau),\ \mathcal{M}_{\text{behav}}(\tau),\ \mathcal{M}_{\text{collab}}(\tau),\ \mathcal{M}_{\text{context}}(\tau)\big),$$

where each $\mathcal{M}_{\text{multi}}, \mathcal{M}_{\text{behav}}, \mathcal{M}_{\text{collab}}, \mathcal{M}_{\text{context}}$ enforces one of the four privacy principles, and h composes their outputs or noise parameters. For example, $\mathcal{M}_{\text{multi}}$ might add calibrated noise to gradient updates at multiple timescales, while $\mathcal{M}_{\text{behav}}$ further imposes policy-level divergence bounds to ensure an entire learned policy does not reveal sensitive information. The function h could be a higher-level controller that orchestrates noise or post-processing across these components, balancing their respective privacy-utility trade-offs.

We note, however, that existing composition theorems for differential privacy are largely tailored to static or i.i.d. settings [6], and directly applying them in sequential RL may be over-restrictive or suboptimal. Investigating novel composition rules that account for temporal dependence, policy-level constraints, and collaborative feedback remains a key open research problem. Future work could explore alternative ways of combining these principles to yield a single privacy budget, or develop contextual composition strategies that selectively apply stricter bounds in high-risk scenarios.

F. Theoretical Bounds on Privacy-Utility Trade-offs

Our Sequential Privacy framework implies that any mechanism satisfying (α, β)–*behavioral privacy* must incur a quantifiable performance cost. Below we state a concrete bound under standard finite-MDP assumptions.

Lemma 4.1 (Privacy–Utility Trade-off (Sketch)): In any finite MDP with discount γ, a mechanism satisfying (α, β)–behavioral privacy (Def. 4.2) must incur

$$\mathbb{E}\big[V^{\pi^*}(s_0) - V^{\pi_{\text{priv}}}(s_0)\big] \ = \ \Omega\big((1-\gamma)/\beta\big).$$

The above results can be obtained by applying the standard performance-difference lemma [52] to the Rényi bound. It mirrors the familiar inverse-scaling trade-offs in differentially-private supervised learning (e.g. $\Omega(1/\epsilon)$ lower bounds in DP-SGD [53]), but here it applies at the level of entire *policies* rather than per-step gradient updates. It concretely demonstrates that *any* mechanism strongly limiting policy divergence must pay a nontrivial price in expected return.

V. Sequential Privacy in Practical Applications

Our Sequential Privacy framework addresses critical challenges in high-stakes domains. Here we demonstrate how the principles can be implemented in three key areas.

A. Healthcare: Privacy-Critical Treatment Optimization

RL systems optimizing treatment strategies must protect temporal patterns in patient care that could reveal both individual conditions and institutional protocols. For chronic conditions like diabetes, blood glucose and insulin adjustment sequences encode sensitive information even when individual decisions are protected. In such clinical settings, policy gradient methods can be adapted for multi-scale privacy:

$$\theta_{t+1} = \theta_t + \eta \cdot \Big(\sum_i \text{clip}\left(\nabla_\theta \log \pi_\theta(a_i|s_i) A(s_i, a_i), C_{\text{med}}\right) + \mathcal{N}(0, \sigma^2_{\text{med}} C^2_{\text{med}} \mathbf{I}) \Big)$$

Here, C_{med} represents the HIPAA-compliant clipping threshold with σ_{med} calibrated to satisfy differential privacy for sensitive medical attributes. Critical in healthcare is adaptive privacy budgeting, where diagnostic phases may receive stronger protection than maintenance phases while still preserving overall temporal pattern privacy.

B. Autonomous Vehicles: Proprietary Strategy Protection

Vehicle fleets generate massive behavioral data encoding navigation strategies and risk assessment algorithms. The behavioral pattern protection principle becomes crucial when companies share driving experiences to enhance safety while protecting proprietary algorithms [45], [46]. Q-learning variants are particularly suited for autonomous vehicle settings, with behavioral pattern privacy implemented through:

$$\pi(a|s) = \frac{\exp(Q(s,a)/\tau_{\text{auto}})}{\sum_{a'} \exp(Q(s,a')/\tau_{\text{auto}})}$$

The temperature parameter τ_{auto} can be dynamically adjusted based on driving context—higher in routine navigation (preserving proprietary algorithms) and lower in safety-critical scenarios where precise behavior is essential. This implementation balances the need to share knowledge of hazardous scenarios while preserving competitive algorithmic advantages.

C. LLMs: Human Feedback Privacy

RLHF systems must protect not only model behavior but also the characteristics of human feedback providers [27], [54]. Even with anonymized instances, preference patterns could reveal annotator demographics through temporal correlations. For collaborative privacy in RLHF, we can implement a stratified protection approach:

$$r_{\text{private}} = r_{\text{original}} + \text{Lap}(\Delta f / \epsilon_{\text{demo}})$$

where demographic-correlated feedback receives stronger protection (ϵ_{demo}) than content-specific feedback. Critically, preference order preservation constraints must be maintained

while obscuring demographic patterns. For federated RLHF settings, secure aggregation with temporal sensitivity weighting can further protect annotator characteristics while preserving useful preference signals.

VI. The Path Forward: A Research Agenda

The preceding sections highlight critical questions for achieving robust privacy in sequential decision-making. Here, we sketch four complementary directions that together form a research agenda for *sequential privacy* in RL.

A. Theoretical Foundations

While classical differential privacy provides a strong baseline in static settings, *sequential* RL poses unique challenges due to overlapping trajectories, temporal dependencies, and adaptive interaction. Researchers must formalize privacy notions specifically for these correlated settings, extending composition theorems to account for overlapping segments or multi-scale observations. For instance, bounding privacy leakage at partial trajectory segments and analyzing privacy amplification under Markovian assumptions remain open problems. Establishing *impossibility* results—where no mechanism can simultaneously achieve strong privacy and high utility for certain classes of RL tasks—would also offer valuable theoretical guidance. Lastly, rigorous empirical metrics are needed to quantify privacy-utility trade-offs *across* different time horizons.

B. Mechanism Design for Temporal Privacy

Existing privacy approaches in RL typically protect either individual timesteps or entire trajectories, risking either excessive noise or unmitigated leakage. Future work must blend *adaptive noise injection* and *multi-scale perturbations* so that data at highly sensitive timesteps is masked more heavily, while allowing enough signal to learn effective policies. Policy-level regularization methods—such as constraining the divergence between learned policies and a reference policy—could further limit the risk of revealing private information through policy behaviors. Additionally, designing lightweight, domain-aware privacy layers for continuous or partially observed environments would expand the applicability of privacy-preserving RL beyond discrete, small-scale benchmarks.

C. Collaborative Privacy Preservation

Federated RL, multi-agent RL, and RLHF settings introduce continuous coordination and real-time feedback among agents or human annotators. Classic individual-level privacy guarantees (e.g., per-user DP) often fail to capture *group-level* or cross-party inferences. New metrics are thus needed to quantify information leakage in group updates or shared gradients, and mechanisms must ensure that aggregated model parameters do not inadvertently reveal *collective* sensitive patterns. In RLHF scenarios, privacy solutions must conceal annotator identities and attributes, even as the model iteratively incorporates feedback to refine its policies. Developing robust defenses against membership inference, gradient inversion, and other adaptive attacks in collaborative RL is paramount for real-world trust.

D. Implementation and Deployment

Bridging theory and practice requires tools for measuring privacy leakage and ensuring scalable algorithmic performance in complex RL tasks. This includes (1) developing standardized benchmarks and simulations that stress-test privacy mechanisms under diverse temporal structures, (2) creating open-source software libraries that integrate privacy-by-design principles into typical RL pipelines (e.g., policy gradient or Q-learning frameworks), and (3) defining domain-specific best practices to satisfy regulatory or ethical constraints in sensitive environments such as healthcare or autonomous driving. Ultimately, practical deployment necessitates reconciling privacy with real-world demands for minimal latency, interpretability, and fault-tolerance, underscoring the need for multi-disciplinary collaboration among ML researchers, domain experts, and policymakers.

VII. Conclusion

Reinforcement learning has rapidly evolved from a research frontier to a technology shaping critical real-world applications in healthcare, transportation, and AI services like language models. Yet existing privacy frameworks, designed primarily for static, pointwise data protection, leave these sequential systems vulnerable. As we have illustrated, privacy breaches in RL can reveal not only isolated data points but entire temporal or behavioral strategies, along with emergent insights about collaborating agents and their contexts.

To address these challenges, we introduce the *Sequential Privacy* framework built on four fundamental principles: multi-scale protection, behavioral pattern protection, collaborative preservation, and context-aware adaptation. Delivering on this vision demands new theory for temporal and group-level privacy, domain-aware mechanisms, and standardized evaluations that balance privacy, utility, and interpretability.

The time is ripe to confront the inseparable link between sequential decision-making and emergent privacy risks. By building on the *Sequential Privacy* principles and open research questions we have posed, the broader AI community can foster a more secure and privacy-preserving foundation for the next generation of reinforcement learning systems.

Acknowledgment

This research is supported by National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative, the Centre for Frontier Artificial Intelligence Research, Institute of High Performance Computing, A*STAR, and the College of Computing and Data Science at Nanyang Technological University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not reflect the views of National Research Foundation, Singapore, and Infocomm Media Development Authority.

REFERENCES

[1] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, 2015.

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[3] G. Dulac-Arnold *et al.*, "Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis," *Machine Learning*, 2021.

[4] N. Corecco, G. Piatti, L. A. Lanzendörfer, F. X. Fan, and R. Wattenhofer, "Suber: An rl environment with simulated human behavior for recommender systems," in *ECAI*, 2024.

[5] X. Lu, F. X. Fan, and T. Wang, "Action and trajectory planning for urban autonomous driving with hierarchical reinforcement learning," *arXiv preprint arXiv:2306.15968*, 2023.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*, 2006.

[7] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[8] N. Papernot, M. Abadi, Úlfar Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *ICLR*, 2017.

[9] M. Gomrokchi, S. Amin, H. Aboutalebi, A. Wong, and D. Precup, "Membership inference attacks against temporally correlated data in deep reinforcement learning," *Ieee Access*, vol. 11, pp. 42 796–42 808, 2023.

[10] P. Ma, Z. Wang, L. Zhang, R. Wang, X. Zou, and T. Yang, "Differentially private reinforcement learning," in *Proceedings of the International Conference on Information and Communications Security*, 2019.

[11] C. J. Cundy, R. Desai, and S. Ermon, "Privacy-constrained policies via mutual information regularized policy gradients," in *AISTATS*, 2024.

[12] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: Information leakage from collaborative deep learning," in *Proceedings of the ACM SIGSAC CCS*, 2017.

[13] X. Fan, Y. Ma, Z. Dai, W. Jing, C. Tan, and B. K. H. Low, "Fault-tolerant federated reinforcement learning with theoretical guarantee," in *Advances in Neural Information Processing Systems*, 2021.

[14] F. X. Fan, Y. Ma, Z. Dai, C. Tan, B. K. H. Low, and R. Wattenhofer, "Fedhql: Federated heterogeneous q-learning," *arXiv preprint arXiv:2301.11135*, 2023.

[15] J. Woo, G. Joshi, and Y. Chi, "The blessing of heterogeneity in federated q-learning: Linear speedup and beyond," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[16] P. Jordan, F. Grötschla, F. X. Fan, and R. Wattenhofer, "Decentralized federated policy gradient with byzantine fault-tolerance and provably fast convergence," in *Proceedings of the 2024 International Conference on Autonomous Agents and Multiagent Systems*, 2024.

[17] S. Yue, X. Hua, Y. Deng, L. Chen, J. Ren, and Y. Zhang, "Momentum-based contextual federated reinforcement learning," *IEEE/ACM Transactions on Networking*, 2024.

[18] Y. Di, H. Shi, R. Ma, H. Gao, Y. Liu, and W. Wang, "Fedrl: A reinforcement learning federated recommender system for efficient communication using reinforcement selector and hypernet generator," *ACM Trans. Recomm. Syst.*, 2024.

[19] J. Woo, L. Shi, G. Joshi, and Y. Chi, "Federated offline reinforcement learning: Collaborative single-policy coverage suffices," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[20] W. Jiang, J. Wang, X. Zhang, W. Bao, C. Tan, and F. X. Fan, "Fedhpd: Heterogeneous federated reinforcement learning via policy distillation," *arXiv preprint arXiv:2502.00870*, 2025.

[21] OpenAI, "Chatgpt," OpenAI Blog, 2023.

[22] DeepSeek-AI, "Deepseek-v3 technical report," https://arxiv.org/abs/2412.19437, 2024, arXiv:2412.19437.

[23] P. F. Christiano, J. Leike, T. Brown *et al.*, "Deep reinforcement learning from human preferences," in *Proceedings of NIPS*, 2017.

[24] N. Ziegler, J. Stiennon, and M. Brundage, "Fine-tuning language models from human preferences," https://arxiv.org/abs/1909.08593, 2019, arXiv:1909.08593.

[25] J. Stiennon *et al.*, "Learning to summarize with human feedback," in *Proceedings of NeurIPS*, 2020.

[26] N. Corecco, G. Piatti, L. A. Lanzendörfer, F. X. Fan, and R. Wattenhofer, "An llm-based recommender system environment," *arXiv e-prints*, 2024.

[27] F. X. Fan, C. Tan, Y.-S. Ong, R. Wattenhofer, and W.-T. Ooi, "Fedrlhf: A convergence-guaranteed federated framework for privacy-preserving and personalized rlhf," *arXiv preprint arXiv:2412.15538*, 2024.

[28] A. von Recum, C. Schnabl, G. Hollbeck, S. Alberti, P. Blinde, and M. von Hagen, "Cannot or should not? automatic analysis of refusal composition in ift/rlhf datasets and refusal behavior of black-box llms," *arXiv preprint arXiv:2412.16974*, 2024.

[29] European Commission, "General data protection regulation (gdpr)," https://gdpr.eu/, accessed: April 2025.

[30] U.S. Department of Health & Human Services, "Health insurance portability and accountability act (hipaa)," https://www.hhs.gov/hipaa/index.html, accessed: April 2025.

[31] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, "Differential privacy under continual observation," in *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, 2010.

[32] C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin, "Pan-private streaming algorithms," in *Proceedings of ICS*, 2010.

[33] J. Tsitsiklis, K. Xu, and Z. Xu, "Private sequential learning," in *Proceedings of COLT*, 2018.

[34] N. Agarwal and K. Singh, "The price of differential privacy for online learning," in *Proceedings of ICML*, 2017.

[35] J. Sakuma, S. Kobayashi, and R. N. Wright, "Privacy-preserving reinforcement learning," in *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[36] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of ACM CCS*, 2015.

[37] R. Gilad-Bachrach *et al.*, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[38] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[39] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2015.

[40] Apple, "Learning with privacy at scale," Apple Machine Learning Journal, 2017.

[41] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," in *Proceedings of IEEE FOCS*, 2018.

[42] Z. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proceedings of NeurIPS*, 2019.

[43] B. Balle, M. Gomrokchi, and D. Precup, "Differentially private policy evaluation," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2130–2138.

[44] X. Zhang, M. M. Khalili, and M. Liu, "Differentially private real-time release of sequential data," *ACM Transactions on Privacy and Security*, 2022.

[45] SAE International, "Taxonomy and definitions for terms related to driving automation systems," SAE International, 2018.

[46] S. Karnouskos and F. Kerschbaum, "Privacy and integrity considerations in hyperconnected autonomous vehicles," *Proceedings of the IEEE*, 2017.

[47] W. Zhao, Y. Sang, N. Xiong, and H. Tian, "Privacy-preserving deep reinforcement learning based on differential privacy," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024.

[48] I. Mironov, "Rényi Differential Privacy," in *IEEE Computer Security Foundations Symposium (CSF)*, 2017.

[49] X. Pan, W. Wang, X. Zhang, B. Li, J. Yi, and D. Song, "How you act tells a lot: Privacy-leaking attack on deep reinforcement learning." in *AAMAS*, 2019.

[50] A. Rajabi, B. Ramasubramanian, A. A. Maruf, and R. Poovendran, "Privacy-preserving reinforcement learning beyond expectation," in *Proceedings of the IEEE Conference on Decision and Control (CDC)*, 2022.

[51] G. Vietri, B. Balle, A. Krishnamurthy, and Z. S. Wu, "Private reinforcement learning with pac and regret guarantees," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[52] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002, pp. 267–274.

[53] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *FOCS*, 2014.

[54] N. Carlini, F. Tramèr *et al.*, "Extracting training data from large language models," in *USENIX Security Symposium*, 2021.