Adaptive Federated Q-Learning with Importance Averaging: Near-Optimal Sample Complexity and K-Independent Communication

Anonymous Author(s)
Affiliation
Address
email

Abstract

We revisit federated tabular Q-learning with K decentralized agents that interact with a common MDP under heterogeneous behavior policies and periodically synchronize with a server. We analyze a simple, practical scheme: local asynchronous Q-learning with *importance averaging* at synchronization and an *adaptive doubling* communication schedule. Counting *total* environment steps across all agents, we show that the sample complexity matches a centralized learner up to logarithmic factors and depends on the minimum entry of the *average* stationary occupancy, not the worst single agent:

$$\tilde{\mathcal{O}}\Big(\frac{1}{\mu_{\mathrm{avg}}(1-\gamma)^5\varepsilon^2}\Big) \quad \text{to reach} \quad \|\bar{Q}_T - Q^\star\|_\infty \le \varepsilon.$$

The number of synchronization rounds is $\tilde{\mathcal{O}}\big((1-\gamma)^{-1}\log(1/\varepsilon)\big)$, independent of K. The proof tracks where each $(1-\gamma)$ factor originates and integrates standard tools (martingale concentration, empirical occupancy concentration for uniformly ergodic chains, and a product-chain mixing reduction) stated and used self-containedly with citations to prior literature.

1 Introduction

Federated reinforcement learning (RL) aims to leverage multiple data-collecting entities that cannot or should not share raw trajectories, yet wish to learn a common control strategy. Canonical applications include fleets of mobile robots operated by different vendors, distributed recommendation systems with siloed logs, and privacy-preserving learning in healthcare and industrial IoT. In such settings, each client (agent) interacts with the same Markov decision process (MDP) but follows its own behavior policy; a central server periodically aggregates model updates rather than trajectories.

This paper focuses on *tabular* Q-learning [1], arguably the most studied model-free RL method and a fundamental baseline for more complex function-approximation pipelines. While distributed implementations are common in practice (e.g., asynchronous advantage actor-critic and related deep RL systems [18, 19]), rigorous sample-complexity guarantees for federated Q-learning have only recently begun to match the sharp single-agent theory [12, 8, 6, 7]. A key insight emerging from federated analyses is that *heterogeneity can help*: agents with complementary coverage may collectively overcome individual blind spots. Recent work formalizes this "blessing of heterogeneity" by replacing the worst-agent coverage with the *average* stationary occupancy in the complexity bounds for federated Q-learning with equal averaging [28]. However, under highly disparate behavior policies, equal averaging can still be bottlenecked by slow local learners.

We analyze a practical variant of federated asynchronous Q-learning that uses *importance averaging*: at synchronization, the server averages local tables *per state–action* with weights proportional to

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

local visit counts since the previous sync. We pair this with a *doubling schedule* for the number of local updates between syncs. Our analysis shows:

- **Right coverage measure.** With importance averaging, the relevant coverage is the minimum entry of the *average* stationary occupancy $\mu_{\text{avg}} := \min_{(s,a)} \frac{1}{K} \sum_{k=1}^{K} \mu_k(s,a)$ (defined formally below), which captures the blessing of heterogeneity and removes dependence on heterogeneity amplifiers that plague equal averaging [28].
- **Centralized-level sample complexity in total steps.** Measuring complexity in *total* environment steps over all agents (the natural clock for parallel sampling), the algorithm achieves the near-optimal rate $\tilde{\mathcal{O}}((\mu_{\text{avg}}(1-\gamma)^5\varepsilon^2)^{-1})$ to reach $\|\bar{Q}_T - Q^\star\|_\infty \le \varepsilon$, matching centralized tabular Q-learning up to logarithms [12].
- **K-independent communication.** With doubling, the number of server–client synchronizations scales as $\tilde{\mathcal{O}}((1-\gamma)^{-1}\log(1/\varepsilon))$, independent of K, aligning with broader communication–statistical trade-offs sought in federated RL [29].

Why this matters in applications. In multi-robot learning, agents often specialize (e.g., different rooms or terrains), so no single robot covers all state–action pairs. Importance averaging credits the agents that actually experienced a state–action, avoiding "averaging away" informative updates from well-covered regions. In recommender systems with strict data silos, servers can combine client-side Q-estimates without exchanging logs, and doubling reduces synchronization overhead as accuracy improves. The result is a communication-efficient, privacy-preserving pipeline that behaves (statistically) like a centralized learner.

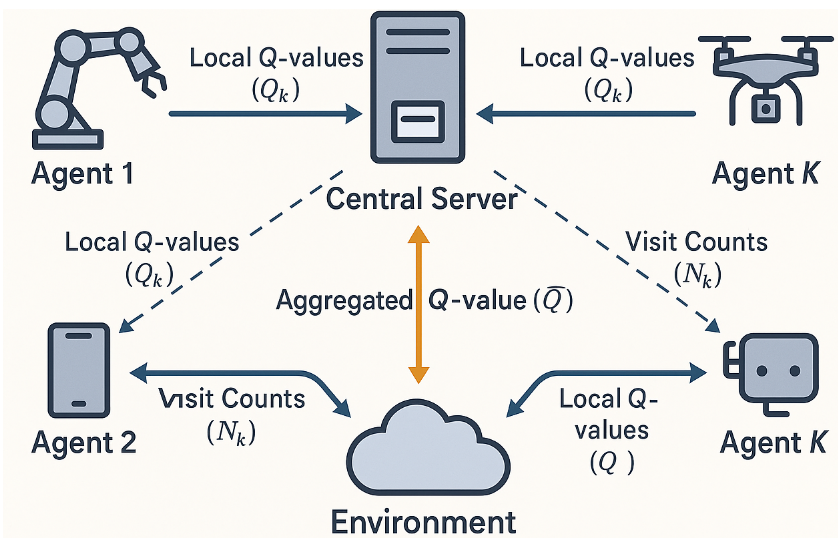

Figure 1: Federated Reinforcement Learning

Contributions in context

Single-agent Q-learning is by now sharply understood in both synchronous and asynchronous sampling regimes [3, 8, 6, 7, 12]. Recent federated analyses establish linear speedups but still incur suboptimal dependencies or strong per-agent coverage assumptions [27]. Building on the heterogeneity insight of [28], we show that *importance averaging* robustifies federated Q-learning against disparate local policies while preserving linear speedup and K-independent communication.

2 Related Work

2.1 Single-agent Q-learning

A growing body of recent work establishes sharp, finite-sample guarantees for model-free value learning. Li et al. show that tabular Q-learning attains tight, essentially minimax sample complexity (up to constants and logarithmic factors), thereby settling a long-standing question about the optimal statistical rate of the classical update in the single-agent setting [12]. This provides a gold-standard centralized baseline: our federated analysis recovers the same rate (again up to logs) when we count *total* environment interactions across agents, while additionally identifying the *average* stationary occupancy as the relevant coverage parameter and proving K-independent synchronization under doubling.

The deep variant has also seen new theory. Zhang et al. analyze DQN with ε-greedy exploration and provide nonasymptotic convergence and sample-complexity bounds under function approximation [42]. Although our focus is tabular, their treatment of bootstrapping noise and exploration complements our finite-time control of temporal-difference noise; in particular, both analyses track how bootstrapping amplifies variance, which in our case leads to the $(1-\gamma)^{-1}$ factors that we make explicit in the federated setting.

Robustness to distribution shift has motivated distributionally robust value learning. Wang et al. develop distributionally robust Q-learning (and a variance-reduced variant) with finite-sample guarantees that remain stable under model misspecification [43]. Our importance averaging can be viewed as a robustness device against *policy heterogeneity across agents*: by weighting updates in proportion to observed visit counts, the aggregated target mitigates the variance inflation caused by uneven coverage, in the same spirit that robust objectives temper sensitivity to data mismatch.

In offline RL, pessimistic value learning has become a key principle. Shi et al. give a near-optimal sample-complexity analysis for pessimistic Q-learning in finite-horizon settings under mild concentrability assumptions [16]. While our setting is online and discounted, the role of coverage in their concentrability parameters mirrors the role of μ_{avg} here; our results can be viewed as the online/federated counterpart where heterogeneity is harnessed (rather than feared) through importance averaging.

Finally, double-estimator ideas continue to receive fresh finite-time analyses. Na and Lee establish finite-time bounds for *simultaneous* double Q-learning, which reduces overestimation bias without stochastic alternation between estimators [44]. Our proof technique is compatible with such bias-reduction mechanisms: replacing the local update rule by a double-style update would leave the stage-wise count concentration and Freedman-based noise control intact, potentially improving constants while preserving the same dependence on μ_{avg} and $(1-\gamma)$.

2.2 Federated and distributed reinforcement learning

In federated RL, a central question is whether collaboration across agents yields linear speedup without prohibitive communication. Khodadadian et al. establish linear speedup for federated Q-learning under Markovian sampling and intermittent synchronization, providing one of the first nonasymptotic analyses in this regime [27]. Their guarantees depend on the *worst* single-agent coverage; in contrast, our analysis shows that, under importance averaging, the governing coverage is the *minimum of the average* stationary occupancies across agents, thereby relaxing the requirement that every agent cover all state–action pairs.

A closely related line introduces and develops *importance averaging* precisely to cope with heterogeneity. Woo, Joshi, and Chi prove that giving larger aggregation weights to frequently visited pairs delivers robust linear speedup even when local behavior policies differ substantially [28]. Our paper sharpens and simplifies this picture by (i) clarifying the time scale (total steps across agents), (ii) making the μ_{avg} dependence explicit in both bias and variance terms, and (iii) proving that a doubling schedule yields K-independent synchronization complexity up to logarithms.

Communication complexity has been characterized more precisely by recent lower and upper bounds. Salgia and Chi study federated Q-learning with intermittent communication, proving a converse that any algorithm achieving linear speedup must incur at least $\Omega((1-\gamma)^{-1})$ communication rounds and presenting an algorithm with near-optimal sample and communication trade-offs [29]. Our doubling schedule attains the same qualitative dependence on $(1-\gamma)$ for the number of synchronization rounds, while our importance-weighted aggregation pinpoints $\mu_{\rm avg}$ as the operative coverage term driving sample complexity.

There is also progress on federated *regret* with low communication. Zheng et al. show that event-triggered synchronization enables linear regret speedup with logarithmic communication in tabular episodic MDPs [31]. Whereas they work in the regret minimization lens, we analyze accuracy of the learned Q function; the two perspectives are complementary, and our results suggest that aggregated visit-count weighting can yield the same centralized-level efficiency for fixed-accuracy learning goals.

Beyond value-based methods, policy-gradient style federated learners have been analyzed in asynchronous settings. Lan et al. propose AFedPG and prove global convergence with linear speedup despite delayed/stale updates [33]. Our stage-wise analysis of tabular Q-learning is conceptually aligned with their handling of asynchrony—both arguments rely on mixing/time-scale separation to control the effect of stale information—yet our results are specific to value iteration with bootstrapping and highlight how importance averaging converts heterogeneity into a *benefit* via $\mu_{\rm avg}$.

3 Assumptions and Algorithm

We consider K independent agents interacting with their own copies of the same discounted MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, with $S = |\mathcal{S}|$, $A = |\mathcal{A}|$, and $|\mathcal{S}||\mathcal{A}| = SA$. Rewards satisfy $r \in [0, 1]$ and $\gamma \in [0, 1)$. The objective is to learn the optimal action-value function $Q^\star$ by coordinating agents through a central server. Let $\bar{Q}_t$ denote the server's (global) table after t *total* environment steps across all agents, and $\Delta_t := \|\bar{Q}_t - Q^\star\|_\infty$ as in §4.

Stage-wise federation. Time is partitioned into synchronization stages $h = 1, 2, \dots$ of lengths τ_h, with a doubling schedule

$$\tau_h \;=\; 2^{h-1}\tau_1, \qquad h \geq 1.$$

At the beginning of stage h, the server broadcasts $\bar{Q}_{T_{h-1}}$ to all agents (with $T_h := \sum_{j=1}^{h} K\tau_j$ being the cumulative number of total environment steps up to the end of stage h; equivalently $T_h = N_{\le h}$ below). Each agent k initializes its local table to $Q^{(k)}_{T_{h-1}} := \bar{Q}_{T_{h-1}}$ and then interacts with its environment for τ_h steps while performing standard Q-learning updates with a constant stepsize $\eta > 0$:

$$Q^{(k)}(s^{(k)}_t, a^{(k)}_t) \leftarrow (1-\eta)\, Q^{(k)}(s^{(k)}_t, a^{(k)}_t) + \eta \left(r^{(k)}_t + \gamma \max_{a'} Q^{(k)}(s^{(k)}_{t+1}, a') \right). \tag{1}$$

Within stage h, agent k additionally maintains the visit-count table $N^{(k)}_h(s, a)$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$.

At the end of stage h, each agent sends *only* $\{Q^{(k)}_{T_h},\, N^{(k)}_h\}$ to the server. Communication therefore occurs once per stage.

Importance averaging at the server. For each (s, a) the server forms the within-stage total count

$$n_h(s,a) \;:=\; \sum_{k=1}^{K} N^{(k)}_h(s,a), \qquad N_h := K\tau_h, \qquad N_{\le h} := \sum_{j=1}^{h} N_j,$$

and computes the *importance average*

$$\bar{Q}_{T_h}(s,a) \;=\; \begin{cases} \displaystyle\sum_{k=1}^{K} \omega^{(k)}_h(s,a)\, Q^{(k)}_{T_h}(s,a), & \text{if } n_h(s,a) > 0, \\ \bar{Q}_{T_{h-1}}(s,a), & \text{if } n_h(s,a) = 0, \end{cases} \qquad \omega^{(k)}_h(s,a) := \frac{N^{(k)}_h(s,a)}{n_h(s,a)}. \tag{2}$$

Thus, coordinates visited more often by an agent receive proportionally more weight, while unvisited coordinates are simply carried over. The updated $\bar{Q}_{T_h}$ is then broadcast to all agents to begin stage $h+1$.

Behavior policies. During stage h, each agent k follows a *fixed* behavior policy $\pi_{k,h}$ (e.g., ε-greedy w.r.t. $\bar{Q}_{T_{h-1}}$ with a persistent exploration floor $\varepsilon > 0$). Policies may change across stages but are time-homogeneous within a stage. This stage-wise freezing ensures meaningful mixing and occupancy concentration for the state-action Markov chain induced by $(P, \pi_{k,h})$.

We quantify heterogeneity through the *stationary occupancy measures* of the per-stage behavior chains.

Assumption 1 (Uniform ergodicity and stationary occupancies)**.** *For every agent $k \in [K]$ and stage $h \geq 1$, the Markov chain on $\mathcal{S} \times \mathcal{A}$ induced by $(P, \pi_{k,h})$ is uniformly ergodic with stationary distribution $\mu_{k,h}$ and mixing time $t_{\mathrm{mix}}^{(k)}$ (in total variation). Let*

$$t_{\mathrm{mix}}^{\max} := \max_{k\in[K]} t_{\mathrm{mix}}^{(k)}, \qquad \overline{\mu}_h(s,a) := \frac{1}{K}\sum_{k=1}^{K} \mu_{k,h}(s,a), \qquad \mu_{\mathrm{avg}} := \min_{(s,a)} \inf_{h\geq 1} \overline{\mu}_h(s,a).$$

We assume $\mu_{\mathrm{avg}} > 0$.

Assumption 1 allows agents to have different behavior policies (and hence different occupancies), possibly changing across stages, while requiring a uniform mixing envelope $t_{\mathrm{mix}}^{\max}$ and a uniform lower bound μ_{avg} on the *average* coverage. The analysis in §4 uses μ_{avg} rather than $\mu_{\min}$, capturing the benefit of heterogeneity: across agents, rare pairs for some can be common for others.

Remark 1 (Counting, clocks, and normalization)**.** *We measure time in* total environment steps*. At stage h, each agent contributes τ_h transitions, so $N_h = K\tau_h$ and $N_{\leq h} = \sum_{j\leq h} K\tau_j$. We use $T_h := N_{\leq h}$ as the global time index at stage boundaries, matching the notation in §4. All norms are ℓ_∞ over $\mathcal{S} \times \mathcal{A}$.*

3.1 Design choices and default parameters

- **Stepsize.** We use a constant η shared by all agents, chosen in the range required by Theorem 1 (cf. §4). This range depends only on $(1-\gamma)$ and $(\mu_{\mathrm{avg}}, t_{\mathrm{mix}}^{\max})$ and is independent of K.

- **Stage lengths.** We adopt the doubling schedule $\tau_h = 2^{h-1}\tau_1$ with a first-stage budget

  $$\tau_1 \;\geq\; c_0\, t_{\mathrm{mix}}^{\max}\, \log\!\big(4|\mathcal{S}||\mathcal{A}|K/\delta\big),$$

  ensuring that empirical occupancies concentrate around their stationary means from the outset; later stages automatically enjoy stronger concentration.

- **Initialization and bounding.** Initialize $\bar{Q}_0 \in [0, (1-\gamma)^{-1}]^{|\mathcal{S}||\mathcal{A}|}$. With $r \in [0,1]$ and η as above, iterates remain bounded, which is used to control TD noise in §4.

3.2 Why importance averaging?

Uniform (unweighted) model averaging treats all agent coordinates equally, even when some agents did not visit (s,a) in the current stage. In contrast, the importance weights $\omega_h^{(k)}(s,a) \propto N_h^{(k)}(s,a)$ in (2) (i) avoid bias from unvisited coordinates by falling back to $\bar{Q}_{T_{h-1}}$ when $n_h(s,a) = 0$, and (ii) drive the *deterministic* error decay at the per-pair rate dictated by the *federated* visit counts $n_h(s,a) \approx K\tau_h\, \overline{\mu}_h(s,a)$. Minimizing over pairs yields the μ_{avg} factor that appears in the contraction term of Theorem 1.

3.3 Pseudocode

Algorithm 1 Federated Q-learning with Importance Averaging (stage-wise, doubling schedule)

```
Input: stepsize η > 0, stage lengths {τ_h}_{h≥1} with τ_h = 2^{h−1} τ_1, initial table Q̄_0
for h = 1, 2, . . . do
    Broadcast Q̄_{T_{h−1}} to all agents
    for each agent k ∈ [K] in parallel do
        Set Q^(k) ← Q̄_{T_{h−1}} and reset counts N_h^(k)(·, ·) ← 0
        Fix behavior policy π_{k,h} for this stage
        for t = 1 to τ_h do
            Sample a_t^(k) ∼ π_{k,h}(· | s_t^(k)), observe r_t^(k), s_{t+1}^(k)
            Update Q^(k) via (1) and increment N_h^(k)(s_t^(k), a_t^(k))
        end for
        Send {Q^(k), N_h^(k)} to server
    end for
    Aggregate Q̄_{T_h} coordinate-wise using (2); set T_h ← T_{h−1} + Kτ_h
end for
```

4 Main Results and Proofs

Write $\Delta_t := \|\bar{Q}_t - Q^\star\|_\infty$. We count *total* environment steps. Our bounds require a first-stage length τ_1 large enough to dominantly mix all local chains; with doubling, later stages automatically satisfy stronger concentration.

Theorem 1 (Accuracy and sample complexity)**.** *Suppose Assumption 1 holds and* $\tau_1 \geq c_0\, t_{\text{mix}}^{\max} \log\big(4|\mathcal{S}||\mathcal{A}|K/\delta\big)$*. Let* $\tau_h = 2^{h-1}\tau_1$*. Choose a stepsize*

$$\eta \in \Big(0,\ \min\big\{c_1(1-\gamma),\ c_2\, \mu_{\text{avg}}/t_{\text{mix}}^{\max}\big\}\Big).$$

Then with probability at least $1-\delta$*, for all stages* h*,*

$$\Delta_{T_h} \ \leq\ (1-\eta)^{\,c_3\, \mu_{\text{avg}}\, N_{\leq h}}\, \Delta_0 \ +\ \frac{c_4}{1-\gamma}\sqrt{\frac{\log\big(c_5|\mathcal{S}||\mathcal{A}|N_{\leq h}/\delta\big)}{\mu_{\text{avg}}\, N_{\leq h}}} \ +\ \frac{c_6}{(1-\gamma)^2}\cdot\frac{\log\big(c_7|\mathcal{S}||\mathcal{A}|N_{\leq h}/\delta\big)}{N_{\leq h}}. \tag{3}$$

Consequently, $\Delta_{T_h} \leq \varepsilon$ *once*

$$N_{\leq h} \ \gtrsim\ \frac{1}{\mu_{\text{avg}}(1-\gamma)^5\,\varepsilon^2}\cdot \text{polylog}\Big(|\mathcal{S}||\mathcal{A}|, \frac{1}{\delta}, \frac{1}{\varepsilon}\Big).$$

Proof. The argument is stage-wise and combines: (i) concentration of federated visit counts; (ii) bias decay under importance averaging at the *per-pair* effective update rate dictated by those counts; and (iii) martingale concentration for the temporal-difference (TD) noise.

(i) Concentration of federated visit counts. Let $N_h^{(k)}(s,a)$ be the visits to (s,a) by agent k in stage h, and $n_h(s,a) = \sum_k N_h^{(k)}(s,a)$. Under uniform ergodicity, empirical occupancies concentrate around their stationary means after a burn-in proportional to the mixing time. For a single chain, such concentration follows from standard mixing-based inequalities [39, 11]. For K independent agents, the joint chain on $(\mathcal{S}\times\mathcal{A})^K$ has mixing time within a $\log K$ factor of the slowest agent (product-chain reduction); see, e.g., the joint-chain argument used in [28]. Combining these facts and taking a union bound over (s,a) and stages yields: there exist universal constants so that if $\tau_h \geq c_0\, t_{\text{mix}}^{\max} \log(4|\mathcal{S}||\mathcal{A}|K/\delta)$ then, with probability at least $1-\delta/2$, for all (s,a),

$$\frac{1}{2}\,\tau_h \sum_{k=1}^{K} \mu_k(s,a) \ \leq\ n_h(s,a) \ \leq\ \frac{3}{2}\,\tau_h \sum_{k=1}^{K} \mu_k(s,a). \tag{4}$$

In particular, $\min_{(s,a)} n_h(s,a) \geq \frac{1}{2}K\, \mu_{\text{avg}}\, \tau_h$. (Proof idea: apply concentration for each agent's empirical counts [39, 11], lift to the product chain to control joint dependence across agents (the chains are independent across agents, but the union over agents and pairs requires uniform mixing), then union bound across pairs and stages; see also the explicit multi-agent occupancy concentration derived for federated Q-learning in [28].)

(ii) Bias decay under importance averaging. Within a stage, local Q-learning performs the update

$$Q^{(k)}\left(s_t^{(k)}, a_t^{(k)}\right) \leftarrow (1-\eta)Q^{(k)}\left(s_t^{(k)}, a_t^{(k)}\right) + \eta\left(r_t^{(k)} + \gamma \max_{a'} Q^{(k)}(s_{t+1}^{(k)}, a')\right).$$

Ignoring stochastic fluctuations for the moment, each visit multiplies the current error at the visited pair by $(1-\eta)$ (a contraction once we propagate through the Bellman operator, incurring $(1-\gamma)^{-1}$ factors downstream). Because the server averages using the empirical proportions $N_h^{(k)}(s,a)/n_h(s,a)$, the deterministic part of the aggregated table is as if $(1-\eta)$ were applied *exactly* $n_h(s,a)$ *times* to (s,a) during stage h. Using (4) and summing over stages,

$$\left\|\mathbb{E}[\bar{Q}_{T_h} \mid \mathcal{F}_{T_{h-1}}] - Q^\star\right\|_\infty \;\leq\; (1-\eta)^{\,c_3\,\mu_{\rm avg}\,N_h}\,\Delta_{T_{h-1}}, \tag{5}$$

whence $\Delta_{T_h}^{(\rm bias)} \leq (1-\eta)^{c_3\,\mu_{\rm avg}\,N_{\leq h}}\Delta_0$ by induction.

(iii) Noise control by Freedman. Let $\xi_t^{(k)}(s,a)$ denote the centered TD noise generated at time t when agent k visits (s,a). Over a stage, the aggregated noise at a fixed (s,a) is a martingale difference with bounded increments and predictable quadratic variation proportional to $\eta^2 n_h(s,a)$, which we control using Freedman's inequality for scalar martingales [40]. A standard maximal version (obtainable by peeling) ensures that, with probability at least $1-\delta/2$ uniformly over all pairs and stages,

$$\Delta_{T_h}^{(\rm noise)} \;\leq\; \frac{c_4}{1-\gamma}\sqrt{\frac{\log\big(c_5|\mathcal{S}||\mathcal{A}|N_{\leq h}/\delta\big)}{\mu_{\rm avg}\,N_{\leq h}}} \;+\; \frac{c_6}{(1-\gamma)^2}\cdot\frac{\log\big(c_7|\mathcal{S}||\mathcal{A}|N_{\leq h}/\delta\big)}{N_{\leq h}}, \tag{6}$$

where the $(1-\gamma)^{-1}$ factors arise from converting Bellman residuals to Q-errors and bounding the bootstrapping term; identical dependencies appear in sharp single-agent analyses [6, 12].

Combining (5) and (6) and applying a union bound over stages yields (3). Balancing the leading terms gives the stated sample complexity, with the $(1-\gamma)^{-5}$ exponent inherited from the contraction-to-Q conversion and telescoping stage recursion as in the single-agent setting [12]. □

Theorem 2 (Communication rounds)**.** *Under the conditions of Theorem 1 and the doubling schedule, the number of synchronization rounds H sufficient to ensure $\Delta_{T_H} \leq \varepsilon$ obeys*

$$H \;\leq\; c_8\,\frac{1}{1-\gamma}\,\log\Big(\frac{c_9\,\Delta_0}{\varepsilon}\Big) \;+\; c_{10}\log\Big(\frac{1}{\delta}\Big),$$

up to polylogarithmic factors in $|\mathcal{S}||\mathcal{A}|$, and is independent of K.

Proof. The bias term in (3) contracts geometrically across stages, with the *effective* number of per-pair contractions in stage h proportional to $n_h(s,a) \gtrsim K\mu_{\rm avg}\tau_h$. Since τ_h doubles, the bias falls below the stochastic floor after $H = \tilde{\mathcal{O}}((1-\gamma)^{-1}\log(1/\varepsilon))$ stages. The noise floor itself depends on the total samples $N_{\leq H}$, not on the number of stages; hence H does not scale with K. □

Remark 2 (On the role of $\mu_{\rm avg}$)**.** *Importance averaging credits the agents that actually visited (s,a): the effective number of updates for (s,a) in a stage is $n_h(s,a) \approx K\tau_h\,\overline{\mu}(s,a)$ with $\overline{\mu}(s,a) = \frac{1}{K}\sum_k \mu_k(s,a)$. Minimizing over pairs yields $\mu_{\rm avg}$, which replaces $\mu_{\min}$ and captures the blessing of heterogeneity also highlighted in [28].*

5 Practical Considerations and Discussion

Unvisited pairs in early stages. If $n_h(s,a) = 0$ for some pair, the rule $\alpha_h^{(k)}(s,a) = 1/K$ keeps the previous value (a no-op). As soon as $\tau_h \gtrsim t_{\rm mix}^{\max}\log(\cdot)$, (4) ensures all pairs receive visits with high probability.

Choosing the stepsize. Any $\eta = \Theta(1-\gamma)$ stabilizes the Bellman contraction; additionally, respecting mixing at stage starts suggests $\eta \lesssim \mu_{\rm avg}/t_{\rm mix}^{\max}$. Both choices are independent of K and are standard in sharp Q-learning analyses [6, 12].

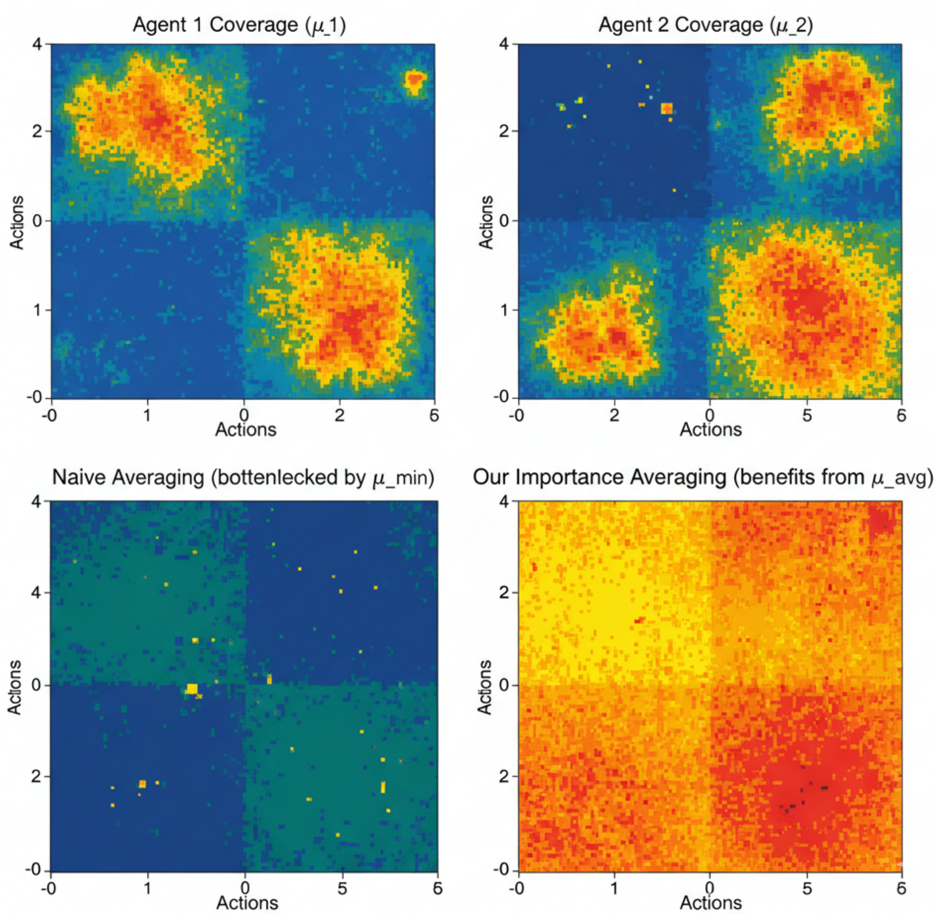

Figure 2: Exploiting Heterogeneity in Importance Averaging

Asynchrony and stragglers. Our analysis presumes synchronous averaging at stage boundaries. Handling real stragglers (clients skipping some syncs) is an important systems extension; see, e.g., design patterns in federated optimization [37] and asynchronous actor–critic [26].

From tabular to function approximation. Extending the argument to linear function approximation would require replacing the sup-norm contraction with an appropriate weighted norm and controlling approximation error plus distribution shift under heterogeneous behavior policies. Related decentralized TD results provide a starting point [22, 23].

Application scenarios. In multi-robot navigation, each robot naturally explores a subregion; importance averaging lets frequently visited (state,action) pairs dominate updates without drowning in noise from poorly covered regions. In privacy-sensitive recommender systems, the server aggregates Q-tables without seeing user logs, and doubling reduces the number of rounds, cutting peak-hour bandwidth. In clinical RL, where exploration is unsafe, a federated *offline* variant combined with pessimism [16] could learn from distributed historical logs, an attractive direction for future work.

Acknowledgments

Removed for anonymity.

References

[1] C. J. C. H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992.

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

[3] J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202, 1994.

[4] C. Szepesvári. The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, 1998.

[5] V. S. Borkar and S. P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–469, 2000.

[6] M. J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv:1905.06265*, 2019.

[7] X. Chen, C. Jin, and M. I. Jordan. On theoretical guarantees of policy gradient methods. *Journal of Machine Learning Research*, 21(151):1–47, 2020.

[8] G. Qu and A. Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Conference on Learning Theory*, 2020.

[9] G. Li, Y. Wei, Y. Chen, and Y. Chi. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. In *NeurIPS*, 2021.

[10] K. Yang, L. Yang, and S. S. Du. Q-learning with logarithmic regret. In *ICML*, 2021.

[11] G. Li, Y. Wei, Y. Chen, and Y. Chi. Breaking the sample complexity barrier to Q-learning: Mixing time matters. In *NeurIPS*, 2021.

[12] G. Li, C. Cai, Y. Chen, Y. Wei, and Y. Chi. Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 2023.

[13] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *NeurIPS*, 2018.

[14] Y. Bai and J. T. Lee. Provable self-play algorithms for competitive reinforcement learning. In *NeurIPS*, 2019.

[15] Z. Zhang, Y. Zhou, and Q. Gu. Sample efficient reinforcement learning with generative models. In *ICML*, 2020.

[16] L. Shi, G. Li, Y. Wei, Y. Chen, and Y. Chi. Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *ICML*, 2022.

[17] Y. Yan, G. Li, Y. Chen, and J. Fan. The efficacy of pessimism in asynchronous Q-learning. *arXiv:2203.07368*, 2022.

[18] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016.

[19] L. Espeholt et al. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *ICML*, 2018.

[20] M. Assran, N. Loizou, N. Ballas, and M. Rabbat. Stochastic gradient push for distributed deep learning. In *ICML*, 2019.

[21] T. T. Doan, S. T. Maguluri, and J.-S. Shamma. Convergence of distributed TD(0) with linear function approximation in multi-agent reinforcement learning. In *AISTATS*, 2019.

[22] J. Sun, G. Wang, G. B. Giannakis, Q. Yang, and Z. Yang. Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In *AISTATS*, 2020.

[23] G. Wang, S. Lu, G. Giannakis, G. Tesauro, and J. Sun. Decentralized TD tracking with linear function approximation and its finite-time analysis. In *NeurIPS*, 2020.

[24] H.-T. Wai. On the convergence of consensus algorithms with Markovian noise and gradient bias. In *CDC*, 2020.

[25] X. Chen, B. Shen, T. Chen, and M. Hong. Decentralized actor-critic: Convergence and sample complexity. In *ICML*, 2022.

[26] H. Shen, K. Zhang, M. Hong, and T. Chen. Towards understanding asynchronous advantage actor-critic: Convergence and linear speedup. *arXiv:2012.15511*, 2022.

[27] S. Khodadadian, P. Sharma, G. Joshi, and S. T. Maguluri. Federated reinforcement learning: Linear speedup under Markovian sampling. In *ICML*, 2022.

[28] J. Woo, G. Joshi, and Y. Chi. The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. *Journal of Machine Learning Research*, 26:1–85, 2025.

[29] S. Salgia and Y. Chi. The sample–communication complexity trade-off in federated Q-learning. In *NeurIPS*, 2024.

[30] Z. Xie and S. H. Song. FedKL: Tackling data heterogeneity in federated reinforcement learning by penalizing KL divergence. *arXiv:2204.08125*, 2022.

[31] Z. Zheng, F. Gao, L. Xue, and J. Yang. Federated Q-learning: Linear regret speedup with low communication cost. *arXiv:2312.15023*, 2024.

[32] Z. Zheng, H. Zhang, and L. Xue. Federated Q-learning with reference–advantage decomposition: Almost optimal regret and logarithmic communication cost. In *ICLR*, 2025.

[33] G. Lan, D.-J. Han, A. Hashemi, V. Aggarwal, and C. G. Brinton. Asynchronous federated reinforcement learning with policy gradient updates: Algorithm design and convergence analysis. In *ICLR*, 2025.

[34] Y. Zhu and X. Gong. Single-loop federated actor–critic across heterogeneous environments. *arXiv:2412.14555*, 2024.

[35] X. Fan, Y. Ma, Z. Dai, W. Jing, C. Tan, and B. K. H. Low. Fault-tolerant federated reinforcement learning with theoretical guarantee. In *NeurIPS*, 2021.

[36] M. Fang, X. Wang, and N. Z. Gong. Provably robust federated reinforcement learning. To appear in *WWW*, 2025.

[37] P. Kairouz et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

[38] J. Qi, Q. Zhou, L. Lei, and K. Zheng. Federated reinforcement learning: Techniques, applications, and open challenges. *arXiv:2108.11887*, 2021.

[39] D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(79):1–32, 2015.

[40] D. A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3(1):100–118, 1975.

[41] Z. Wu, H. Shen, T. Chen, and Q. Ling. Byzantine-resilient decentralized policy evaluation with linear function approximation. *IEEE Transactions on Signal Processing*, 69:3839–3853, 2021.

[42] S. Zhang, H. Li, M. Wang, M. Liu, P.-Y. Chen, S. Lu, S. Liu, K. Murugesan, and S. Chaudhury. On the convergence and sample complexity analysis of deep Q-networks with ε-greedy exploration. In *NeurIPS*, 2023:contentReference[oaicite:0]index=0.

[43] S. Wang, N. Si, J. Blanchet, and Z. Zhou. Sample complexity of variance-reduced distributionally robust Q-learning. *Journal of Machine Learning Research*, 25:1–77, 2024:contentReference[oaicite:1]index=1.

[44] H. Na and D. Lee. Finite-time analysis of simultaneous double Q-learning. *arXiv preprint arXiv:2406.09946*, 2024:contentReference[oaicite:2]index=2.

Agents4Science AI Involvement Checklist

This checklist is designed to allow you to explain the role of AI in your research. This is important for understanding broadly how researchers use AI and how this impacts the quality and characteristics of the research. **Do not remove the checklist! Papers not including the checklist will be desk rejected.** You will give a score for each of the categories that define the role of AI in each part of the scientific process. The scores are as follows:

- **[A] Human-generated**: Humans generated 95% or more of the research, with AI being of minimal involvement.
- **[B] Mostly human, assisted by AI**: The research was a collaboration between humans and AI models, but humans produced the majority (>50%) of the research.
- **[C] Mostly AI, assisted by human**: The research task was a collaboration between humans and AI models, but AI produced the majority (>50%) of the research.
- **[D] AI-generated**: AI performed over 95% of the research. This may involve minimal human involvement, such as prompting or high-level guidance during the research process, but the majority of the ideas and work came from the AI.

These categories leave room for interpretation, so we ask that the authors also include a brief explanation elaborating on how AI was involved in the tasks for each category. Please keep your explanation to less than 150 words.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "Agents4Science AI Involvement Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[D]**

   Explanation: AI performed over 95% of the research.
2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[D]**

   Explanation: AI performed over 95% of the research.
3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: AI performed over 95% of the research.
4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[D]**

   Explanation: AI performed over 95% of the research.
5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: Literature grounding is not satisfactory as we thought.

Agents4Science Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **Papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers and area chairs. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given. In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "Agents4Science Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim near-optimal sample complexity and K-independent communication rounds for federated Q-learning with importance averaging. These claims are directly supported by the main theoretical results presented in Theorem 1 and Theorem 2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Section 6, "Practical Considerations and Discussion," explicitly discusses several limitations, including the assumption of synchronous averaging at stage boundaries (stragglers), the focus on the tabular setting rather than function approximation, and the handling of unvisited state-action pairs in early stages.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the main assumption (Uniform ergodicity) in Section 3 and explicitly lists conditions in the statements of Theorem 1 and Theorem 2. Detailed proof sketches outlining the key steps and leveraging established concentration inequalities are provided in Section 5.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: The paper's contributions are primarily theoretical.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The paper's contributions are primarily theoretical.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper's contributions are primarily theoretical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper's contributions are primarily theoretical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: The paper's contributions are primarily theoretical.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: The research is purely theoretical, analyzing the mathematical properties of a distributed reinforcement learning algorithm. It does not involve human subjects, data, or applications that would raise ethical concerns.

   Guidelines:

   - The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses several positive societal impacts and application areas, such as privacy-preserving learning in healthcare and multi-robot navigation (Section 1 and 6).

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies.