Sheffield Hallam University

Learning while Sleeping: Integrating Sleep-Inspired Consolidation with Human Feedback Learning.

TARAKLI, Imene and DI NUOVO, Alessandro <http://orcid.org/0000-0003-2677-2650>

Available from Sheffield Hallam University Research Archive (SHURA) at:

https://shura.shu.ac.uk/33782/

This document is the Accepted Version [AM]

Citation:

TARAKLI, Imene and DI NUOVO, Alessandro (2024). Learning while Sleeping: Integrating Sleep-Inspired Consolidation with Human Feedback Learning. In: 2024 IEEE International Conference on Development and Learning (ICDL). IEEE. [Book Section]

Copyright and re-use policy

See http://shura.shu.ac.uk/information.html

Sheffield Hallam University Research Archive
http://shura.shu.ac.uk

Learning while Sleeping: Integrating Sleep-Inspired Consolidation with Human Feedback Learning

Imene Tarakli
Sheffield Hallam University
Sheffield, United Kingdom
i.tarakli@shu.ac.uk

Alessandro Di Nuovo
Sheffield Hallam University
Sheffield, United Kingdom
a.dinuovo@shu.ac.uk

***Abstract*—Sleep plays a vital role in developmental learning. It allows the brain to consolidate daily learning experiences by replaying the memories accumulated throughout the day. In this work, we take inspiration from sleep and propose the Inverse Forward Offline Reinforcement Model (INFORM), a scalable framework that first learns a set of behaviours from human evaluative feedback, then consolidates the learning by applying an offline inverse reinforcement learning to the memorised trajectories. Experimental results demonstrate that INFORM is a feedback-efficient method that effectively learns an optimal policy that aligns with the intended behaviour of the human. A comparative analysis shows that the learnt policies are robust to dynamic changes in the environment and the recovered rewards allow the robot to be autonomous in its learning. Project website: https://sites.google.com/view/inform-framework**

***Index Terms*—Developmental Robotics, Cognitive Robotics, Interactive Agents**

I. Introduction

From the joyous cheers greeting a baby's first steps to the reprimand following a misconduct, or the beaming pride after an achievement, humans are consistently exposed to evaluative feedback throughout their lives. This feedback is fundamental for learning, as it helps individuals acquire new skills, make informed decisions, and adapt to changes in the environment [1].

Transferring this ability to learn from evaluative feedback to robots is a natural transition toward smoother Human-Robot Interaction (HRI). By emulating natural human interaction, the teaching process becomes more intuitive and effective, enabling individuals to guide robots using feedback in the same manner they would with their peers. As the feedback precisely tailors the robot's learning, it allows users to easily personalise the robot's behaviour to their desired preferences without needing specific technical skills.

Several studies investigated how to include human feedback within the decision-making process of a robot, framing this approach as Interactive Reinforcement Learning (RL). This framework has proven effective in various real-world applications, enabling users to shape the behaviour of robots by providing feedback on each action of the robots. However, humans do not consider evaluative feedback as a reinforcement per se, but as a means to communicate the correctness of an action, often favouring positive over negative feedback [2], [3]. This way of teaching creates positive-reward cycles – the agent will repeatedly visit the same states to maximise the long-term reward. Consequently, Interactive RL tends to focus on myopic learning - the robot privileges immediate rewards over future ones when making decisions. While this approach accelerates learning by reducing the time spent exploring future rewards, it can limit the robot's understanding of the broader goal of the task, potentially affecting the generalisation and robustness of its performance [4].

In contrast, humans effectively learn from evaluative feedback and are capable of generalising and transferring the knowledge to more complex tasks. However, this learning is not straightforward. Humans do not acquire robust decision-making skills directly from evaluative feedback; it involves various steps, with sleep playing a crucial role [5]. Indeed, studies revealed that deep sleep enhances learning by enabling the consolidation and generalisation of knowledge [6]. Newly acquired skills are initially stored in the short-term memory within the hippocampus. During sleep, these memories are processed and replayed, training additional neurons in the cortex. This process maximises the information extraction from each episode and stores it in the brain's long-term memory, facilitating the generalisation of learning [7].

Moreover, humans do not solely rely on human feedback when learning. they infer the goal and intent of the teacher based on that feedback and internalise the reward to enable further learning on their own [8].

In this paper, we introduce **IN**verse **F**orward **O**ffline **R**einforcement **M**odel (INFORM), a model for learning from evaluative feedback that closely emulates how humans learn. INFORM initially learns from evaluative feedback by using a myopic forward interactive RL to predict the teacher's preferred actions, storing all trajectories in a replay buffer. It then initiates a "sleep phase" by replaying these learning

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955778.

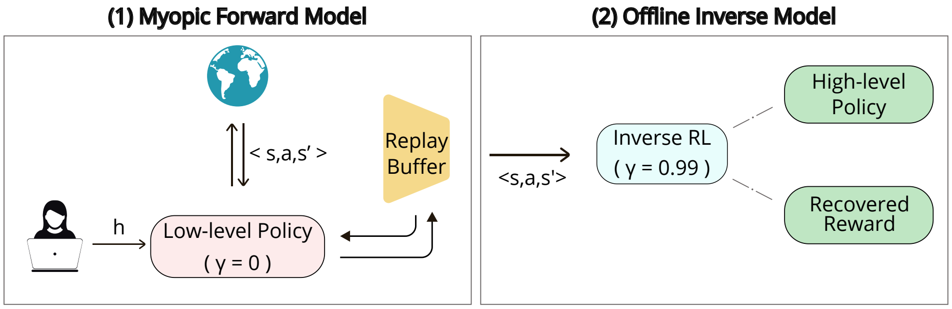


Fig. 1: Illustration of the framework. (1) The agent first learns a low-level policy with a myopic interactive RL. All trajectories of the interaction are stored in a buffer. (2) An offline inverse RL is then applied to the stored trajectories to recover a reward and a policy that better encodes the high-level information of the task.

experiences in an offline mode. During this phase, Inverse RL is applied with a high discount factor, allowing the model to internalise the dynamics of the environment and thus learn a robust policy and reward function from past experiences [9].

Our contribution can be summarised as follows:

- We introduce INFORM, a framework that combines interactive and inverse RL to enable scalable learning from human feedback.
- We report that INFORM is the first successful instance of learning a non-myopic policy from human feedback in a model-free setting for discrete and continuous environments.
- We empirically prove that INFORM is robust to the change of dynamics and distributional shifts in a diverse set of environments.
- We show that the reward function recovered by INFORM effectively captures the high-level goal of the task, enabling autonomous learning.

II. Preliminaries

In this section, we review the relevant definition and notations that are used in the rest of this work.

Reinforcement Learning. RL is a subset of machine learning that aims to solve problems modeled as Markov Decision Processes (MDPs) [10]. An MDP is defined as a tuple $< S, A, T, R, \gamma >$ where S and A are respectively the set of possible states and actions; $T : S \times A \rightarrow S$ is the transition function that gives the probability of transitioning to another state given the actual state and action; $R : S \times A \rightarrow \mathbb{R}$ is the reward function which defines the received reward when performing an action in a state, and $\gamma \in [0, 1]$ is the discount factor that determines the sensitivity of the agent to future decisions. RL aims to learn policies $\pi : S \rightarrow A$ that improve the discounted sum of returns over time. The return of each $< s, a >$ pair is given by the action-value defined as: $Q(s, a) = E_\pi [\sum_{t=0}^{\infty} \gamma^t R(s, a)]$. Policies that maximise the returns rewards are called optimal and are denoted by π^*.

Inverse Reinforcement Learning. In IRL, the problem is modeled as an MDP/R tuple - MDP without any reward function R [11]. Instead, we are given a set of trajectories $\mathcal{D} = \{(s_1, a_1), ..., (s_n, a_n)\}$ that are assumed to be samples from an expert policy π_E. The goal is to find a reward function R that would make the agent reproduce the same behaviour as the expert.

Myopic learning. It occurs when an agent prioritises immediate rewards without considering the long-term consequences of its actions. This approach may lead to suboptimal performance in complex environments as the agent may not account for the potential future rewards or penalties. In RL, a learning approach is considered myopic if it uses a low discount factor (typically, $\gamma <= 0.9$) [4].

III. Related works

Learning from human evaluative feedback. In interactive RL, an agent improves its policy based on the scalar feedback provided by a human teacher. This feedback, interpreted as either a value [12]–[14] or advantage function [15], [16], is non-stationary and inconsistent, making it distinct from standard environmental rewards. Studies found that humans provide more positive than negative feedback, resulting in positive-reward cycles [2]–[4]. This is addressed by myopic learning; however, this approach makes the learning less generalisable. To address this, Knox et al. [4] developed VI-TAMER, a non-myopic model that merges TAMER with a Value-Iteration algorithm for robust learning. While this framework improved the generalisation of the performance, it can only be used in discrete environments with a known transition model, limiting its broader applicability.

Sleep Inspired Reinforcement Learning Massi et al. [17] proposed a neuroscience-inspired RL model incorporating a hippocampal replay mechanism, which demonstrated faster and more efficient learning through a replay buffer developed based on neuroscience knowledge about the hippocampus. Similarly, Tirumala et all. [18] proposed Replay across Experiments (RaE), a framework that reutilises experiences from past experiments, enhancing exploration and accelerating learning by replaying diverse trajectories. While these studies showcase the value of hippocampal-inspired offline learning, they focus on scenarios with well-defined reward functions. Our research, in contrast, focuses on learning derived from human feedback, exploring a new aspect of consolidation learning.

IV. The INFORM Framework

In this section, we present the INverse Forward Offline Reinforcement Model (INFORM), a framework that scalably learns generalised policies and reward functions from human feedback. As depicted in Fig. 1 , the model consists of two phases :

(1) **A Forward model:** In this initial phase, we use a myopic interactive RL based on human feedback to train a preliminary, low-level policy.

(2) **An Offline Inverse model:** Subsequently, we revisit all the trajectories generated by the previous phase and apply a non-myopic offline IRL to derive a policy and reward function that more accurately captures the task's high-level objectives.

A. The Forward Model

The forward model allows a human trainer to tailor the policy of a robot to a specific behaviour. Initially, the robot follows a random policy. The human then provides binary feedback, denoted as f, to assess the correctness of each $<\text{state}, \text{action}>$ pair, and guide the policy update H of the robot toward the desired behaviour by optimising the learning objective:

$$\mathcal{L}\left(\theta_H\right)=\mathbb{E}_{(s,a,H)\sim\mathcal{D}}\left\|\hat{H}\left(s,a;\theta_H\right)-H(s,a)\right\|^2 \qquad (1)$$

During this phase, we use the TAMER model, a widely used framework that effectively learns from evaluative feedback [12].

At the end of each episode, the robot assesses the success of its performance based on the environmental return, which is derived from observations rather than the traditional reward function. It then stores the trajectories with a success or failure tag in a replay buffer for use in the next phase.

B. The Offline Inverse Model

Although the forward model efficiently learns the intended behaviour of humans from evaluative feedback, the resulting policy might not generalise well due to the short-slightness of fully myopic learning with a high discount factor ($\gamma = 0$). In this setting, the robot only cares about immediate actions and relies entirely on human feedback which results in immediate, short-term learning that can limit the robot's understanding of the environment and the task.

To overcome this limitation, we propose to emulate the sleep phase which allows the brain to consolidate the learning by sorting and reinforcing newly encoded memories and transition them into the more abstract, generalised type of memory by training additional neural network [19].

This mechanism is built in the same way as inverse RL, in which policies are derived from given trajectories. Specifically, we take an interest in IQ-learn [9], a dynamic-aware imitation learning model that effectively learns a Q-function from a few demonstrations and uses it to derive both a policy and a reward function. While successful results were obtained in different environments, the method relies on expert demonstrations and, in more intricate continuous environments, necessitates direct interaction to capture dynamic information about the environments.

However, sleep occurs offline without access to optimal expert trajectories. To align inverse learning with this process, we modify the IQ-learn objective function to learn from successful trajectories of the forward model rather than expert ones. Additionally, we substitute online interaction with the environment, with samples from the replay buffer of the first phase, regrouping both successful and unsuccessful trajectories.

$$\max_{Q\in\Omega}\mathcal{J}^*(Q)=-\mathbb{E}_{(s,a,s')\sim\text{success}}\left[Q(s,a)-\gamma V^*\left(s'\right)\right] \\ -\mathbb{E}_{(s,a,s')\sim\text{ replay}}\left[V^\pi(s)-\gamma V^\pi\left(s'\right)\right] \qquad (2)$$

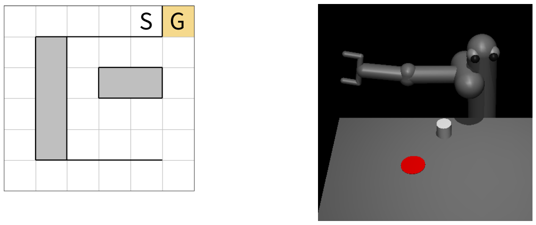
with $V^*(s)=\log\sum_a\exp Q(s,a)$.

From the learnt policy, a reward function, encoding the high-level goal of the task and enabling autonomous learning, can be recovered as follows:

$$r\left(s,a,s'\right)=Q(s,a)-\gamma V^\pi\left(s'\right) \qquad (3)$$

Algorithm 1 Pseudocode of the INFORM Framework

```
=== Forward Model ===
Initialise parameters of $H_\theta$, $\pi_\phi$, and a replay buffer $\mathcal{D}$
for each episode do
    Initialize a temporary buffer $\mathcal{T}$
    for each step do
        $s \leftarrow$ observation
        Select action $a$:
                (Q-learning) $a = argmax(H_\theta(s))$
                (actor-critic) $a = \pi(\cdot \mid s_t; \phi)$
        $s' \leftarrow$ executing a
        $h \leftarrow$ human feedback
        Store in temporary buffer $\mathcal{T} \leftarrow \mathcal{T} \cup (s, a, s', h)$
        Perform a gradient step for $\theta_H$ using Equation 1
        (only with actor-critic) Perform a gradient step for $\phi_\pi$
        using the following equation:
          $\mathcal{L}(\phi_\pi) = \mathbb{E}_{(s,a,\pi)\sim\mathcal{D}}[H(s,a) - \log\pi(a|s;\phi)]$
    end for
    Assess success of trajectories    ▷ $success \in \{0$ or $1\}$
    Store labeled trajectories in replay buffer $\mathcal{D} \leftarrow \mathcal{D} \cup$
    $(\mathcal{T}, success)$
end for
 === Inverse Model ===
Initialise parameters of $Q_\theta$ and optionally $\pi_\phi$
for each step do
    Get batch from replay buffer $\sim \mathcal{D}[(s, a, s', success)]$
    Perform a gradient step for $\theta_Q$ using Equation 2
    (only with actor-critic) Perform a gradient step for $\phi_\pi$
end for
Recover reward function using Equation 3
```



(a) Gridworld (b) Pusher

Fig. 2: Griwrold and Pusher-V4 environments are used to assess the performance of INFORM.

V. Methodology

We structure our methodology to address the following research questions:

- Can INFORM effectively learn a high-level policy from past experiences?
- How robust is the consolidated policy learned by INFORM?
- Is the reward function retrieved by INFORM in alignment with the teacher's intended outcomes?

A. Environments

We evaluate INFORM in two distinct scenarios: a discrete task and a continuous task.

In the discrete setting, we employ the 30-cell gridworld maze (Figure 2a) from [4]. The agent's objective is to navigate from the 'S' cell to the 'G' cell in as few steps as possible, choosing from four directional actions: up, down, left, or right. A trajectory is deemed successful if the goal is reached in under 30 steps. This environment, while straightforward, provides a comprehensive platform for a detailed and observable evaluation of our framework.

For the continuous task, we use the Pusher-V4 task from Mujoco Gymnasium [20], where a robotic arm with multiple joints aims to move a cylinder to a target position using its end effector (Figure 2b). In this environment, a trajectory is considered as successful if the cylinder is within 0.08 units from the target position. The increased complexity of this task, compared to the gridworld maze, enables a more thorough assessment of the performance of INFORM in complex scenarios.

B. Simulated Feedback

This study aims to enhance learning from evaluative feedback, not through direct improvements from the feedback, but by consolidating knowledge derived from human interactions.

For the direct learning from evaluative feedback, we apply established models that have been previously validated with human participants [12]. Therefore, to simplify the experimental process and reduce reliance on human feedback, we implement an oracle to simulate evaluative feedback, facilitating a comprehensive assessment of the INFORM framework.

Similar to Zhang et al. [21], we use a fully trained model as the oracle. For each given state, s, the learning agent selects an action, a, based on its current policy while the oracle simultaneously selects an action, a*, based on its optimal policy. The oracle is then used to calculate the state-action value for both actions. When the learning agent's action produces a Q-value close to Q(s,a*), t it is considered a successful action, leading to positive feedback. In cases where Q-values significantly differ, no feedback is given. This design aims to replicate the positive feedback bias observed in human interactions [22].

$$F(s,a) = \begin{cases} +1 & \text{if } Q(s,a) \geq \alpha Q\left(s,a^{*}\right) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where α is a variable that increases over time to modelize the diminishing return of human evaluative feedback.

C. Implementation

In our implementation of INFORM, we use a Deep Q-Network (DQN) [23] for the discrete gridworld environment and a Soft Actor-Critic (SAC) [24] for the continuous task. We adapted the CleanRL codebase [25] to align with Algorithm 1, and selected hyperparameters based on the recommendations of Garg et al. [9]. Detailed implementation specifics are available on the project's website.

VI. Results & Discussion

A. Consolidating learning from human feedback

We assessed whether INFORM could effectively consolidate learning from human feedback by developing a non-myopic, high-level offline policy from previous experiences, without additional interaction with the environment.

We evaluated the framework on the two environments: gridworld and Pusher-V4. For both tasks, we first learnt a low-level policy from the oracle feedback using TAMER. INFORM then consolidated the myopic learning by applying offline inverse RL to the past trajectories with a high-discount factor.

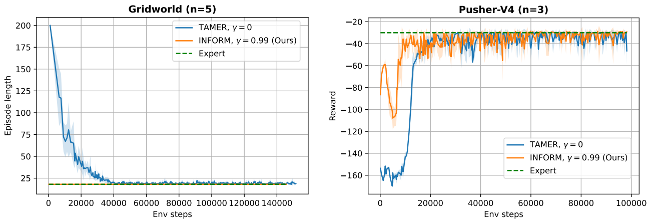
Figure 5 shows the obtained results. INFORM (orange line) rapidly aligns with the performance of TAMER, representing the myopic policy derived from human feedback (blue line), and matches the expert performance (dashed line).

This demonstrates that the learning objective of INFORM successfully integrates learning from both successful and unsuccessful trajectories to consolidate knowledge by learning a high-level policy for each task.

B. Robustness to dynamic changes

We assessed whether the high-level policy obtained with INFORM is robust against changes in environmental dynamics.

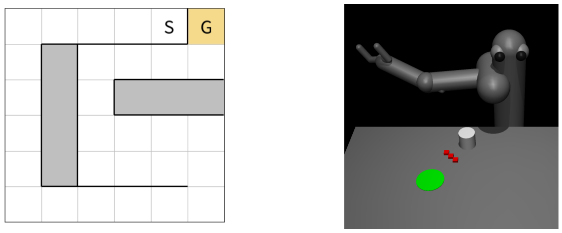
Following the methodology of Eysenbach et al. [26], we modified the Pusher-v4 task by introducing an obstacle along the perpendicular bisector between the puck's initial position and the goal. This obstacle comprises three axis-aligned blocks, each 3 cm wide, located at (0.32, -0.2), (0.35, -0.23), and (0.38, -0.26). The modified environment is illustrated in figure 4b.



(a) Gridworld

(b) Pusher

Fig. 3: Evaluation of INFORM and TAMER across discrete and continuous Tasks. The expert performance (green dashed line) sets the target for each environment. Higher values indicate superior performance in Pusher, while lower values are better in Gridworld. INFORM effectively recovers non-myopic policies, matching TAMER's performance.



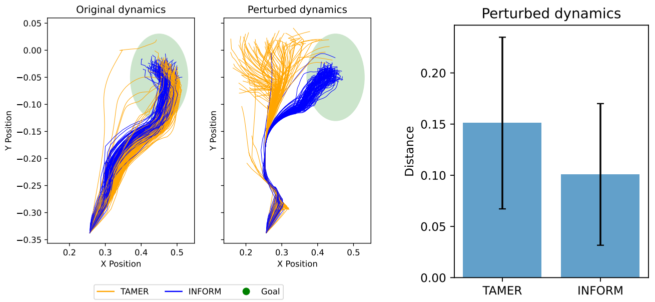
(a)

(b)

Fig. 4: Perturbation in environmental dynamics. (a) A wall is added in Gridworld to block the optimal path. (b) The optimal trjaectory is obstructed with an obstacle.

We initially trained TAMER and INFORM in the original environment and then tested them in the environment with altered dynamics over 100 episodes each, for three runs. To ensure consistency, the same initial state was used in all episodes. The robustness of each model was evaluated based on the distance between the final position of the object and the target's position.

Figure 5b depicts the success rate of both models in the original and perturbed dynamics. We notice the high-level policy obtained with INFORM significantly moves the object to its target closer than the low-level policy obtained with TAMER, $p < 0.0001$, despite the presence of the obstacle. As observed in figure 5a, the policy obtained with INFORM is more likely to go around the obstacles and reach the goal than TAMER. We posit that the high discount factor enables the model a more accurate representation of the dynamics of the environment, which can contribute to the robustness of the model. In contrast, a myopic policy, primarily oriented towards the optimal trajectories, may underperform when deviating from these trajectories. However, it's important to acknowledge that despite INFORM's robust policy, there was significant variance in performance across various seeds. This variability suggests that INFORM could be further improved to enhance the consistency of the results.



(a)

(b)

Fig. 5: Robustness evaluation. (a) Comparison of TAMER and INFORM trajectories on both original and perturbed environments. (b) Comparison of distance from target over 3 runs. INFORM significantly moves the object closer to the goal compared to TAMER, learning a high-level policy that more effectively navigates around obstacles.

C. Reward evaluation

In this experiment, we aimed to evaluate the reward function recovered through INFORM. For that, we modified the training environment to render the policy learned with human feedback suboptimal. This modification ensures that the reward function recovered by INFORM captures the high-level goal of the task, rather than merely mimicking optimal behaviour. Similarly to [4], we modify the gridworld environment by blocking the optimal path with a wall (Fig. 4a). We first trained INFORM in the original gridworld and recovered a reward function using equation. Using this recovered reward function, we trained new agents in the modified environment, without any human feedback. These agents were trained using tabular Q-learning across 300 episodes for ease of implementation. To compare the performances, we train similar agents using the environmental reward (+1 for reaching the goal, 0 otherwise).

Figure 6 illustrates the learning performance in this modified environment. The results show that the agents learning with INFORM reward effectively converge to the optimal policy, paralleling expert performance, and more rapidly than agents learning with the environment reward. Further analysis of the reward function revealed the one recovered from INFORM is denser than the environmental reward, providing agents with more information about the environment, which allows a faster convergence of the policy.

In contrast, when evaluating TAMER-trained policies in this modified setting (without further training), it was observed that they uniformly failed to achieve the goal. These policies, developed from direct human feedback, did not incorporate adjustments for the new wall, requiring additional human guidance to adapt to these environmental changes.

VII. Conclusion

In this study, we presented a sleep-inspired framework that consolidates learning from human feedback. Initially, our model learns a myopic policy through human feedback,

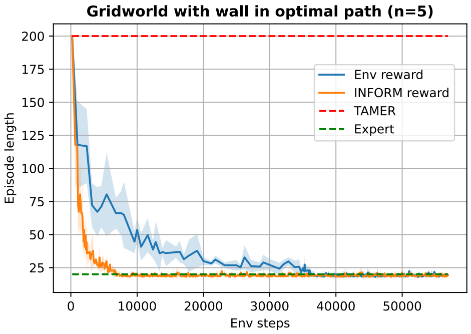


Fig. 6: Learning curve of agents in the gridworld with a wall blocking the optimal path. Agents learning with INFORM reward successfully and rapidly learn an optimal policy

then employs offline inverse reinforcement learning (RL) on past experiences to simulate sleep consolidation and enhance learning.

Our comparative evaluation of benchmark RL environments demonstrated that INFORM effectively acquires high-level optimal policies in both discrete and continuous tasks, showcasing its robustness against dynamic changes. Moreover, the reward function retrieved by INFORM aligns with the high-level objectives of tasks, enabling scalable autonomous learning.

This research lays the groundwork for future exploration in consolidation learning from human feedback. Subsequent studies could validate the framework further by testing live human-robot interaction and examining the influence of human irrationality in providing feedback on INFORM's effectiveness. Furthermore, future works should investigate the impact of the quality and amount of data available during learning with humans for the consolidation to be effective.

Acknowledgement

We would like to thank Sarah Hamburg for her invaluable neuroscience insights and for reviewing an earlier version of this manuscript.

References

[1] B. Wisniewski, K. Zierer, and J. Hattie, "The power of feedback revisited: A meta-analysis of educational feedback research," *Frontiers in Psychology*, vol. 10, p. 3087, 2020.

[2] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, vol. 172, no. 6-7, pp. 716–737, 2008.

[3] M. K. Ho, F. Cushman, M. L. Littman, and J. L. Austerweil, "People teach with rewards and punishments as communication, not reinforcements." *Journal of Experimental Psychology: General*, vol. 148, no. 3, p. 520, 2019.

[4] W. B. Knox and P. Stone, "Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance," *Artificial Intelligence*, vol. 225, pp. 24–50, 2015.

[5] R. Stickgold, "Sleep-dependent memory consolidation," *Nature*, vol. 437, no. 7063, pp. 1272–1278, 2005.

[6] R. Huber, M. Felice Ghilardi, M. Massimini, and G. Tononi, "Local sleep and learning," *Nature*, vol. 430, no. 6995, pp. 78–81, 2004.

[7] D. S. Ramanathan, T. Gulati, and K. Ganguly, "Sleep-dependent reactivation of ensembles in motor cortex promotes skill consolidation," *PLoS biology*, vol. 13, no. 9, p. e1002263, 2015.

[8] M. K. Ho and T. L. Griffiths, "Cognitive science as a source of forward and inverse models of human decisions for robotics and control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 33–53, 2022.

[9] D. Garg, S. Chakraborty, C. Cundy, J. Song, and S. Ermon, "Iq-learn: Inverse soft-q learning for imitation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4028–4039, 2021.

[10] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[11] A. Y. Ng, S. Russell *et al.*, "Algorithms for inverse reinforcement learning." in *Icml*, vol. 1, 2000, p. 2.

[12] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The tamer framework," in *Proceedings of the fifth international conference on Knowledge capture*, 2009, pp. 9–16.

[13] A. Najar, O. Sigaud, and M. Chetouani, "Training a robot with evaluative feedback and unlabeled guidance signals," in *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 2016, pp. 261–266.

[14] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone, "Deep tamer: Interactive agent shaping in high-dimensional state spaces," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[15] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman, "Interactive learning from policy-dependent human feedback," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2285–2294.

[16] D. Arumugam, J. K. Lee, S. Saskin, and M. L. Littman, "Deep reinforcement learning from policy-dependent human feedback," *arXiv preprint arXiv:1902.04257*, 2019.

[17] E. Massi, J. Barthélemy, J. Mailly, R. Dromnelle, J. Canitrot, E. Poniatowski, B. Girard, and M. Khamassi, "Model-based and model-free replay mechanisms for reinforcement learning in neurorobotics," *Frontiers in Neurorobotics*, vol. 16, p. 864380, 2022.

[18] D. Tirumala, T. Lampe, J. E. Chen, T. Haarnoja, S. Huang, G. Lever, B. Moran, T. Hertweck, L. Hasenclever, M. Riedmiller *et al.*, "Replay across experiments: A natural extension of off-policy rl," *arXiv preprint arXiv:2311.15951*, 2023.

[19] S. Brodt, M. Inostroza, N. Niethard, and J. Born, "Sleep—a brain-state serving systems memory consolidation," *Neuron*, vol. 111, no. 7, pp. 1050–1075, 2023.

[20] M. Towers, J. K. Terry, A. Kwiatkowski, J. U. Balis, G. d. Cola, T. Deleu, M. Goulão, A. Kallinteris, A. KG, M. Krimmel, R. Perez-Vicente, A. Pierré, S. Schulhoff, J. J. Tai, A. T. J. Shen, and O. G. Younis, "Gymnasium," Mar. 2023. [Online]. Available: https://zenodo.org/record/8127025

[21] R. Zhang, D. Bansal, Y. Hao, A. Hiranaka, J. Gao, C. Wang, R. Martín-Martín, L. Fei-Fei, and J. Wu, "A dual representation framework for robot learning with human guidance," in *Conference on Robot Learning*. PMLR, 2023, pp. 738–750.

[22] M. K. Ho, J. MacGlashan, M. L. Littman, and F. Cushman, "Social is special: A normative framework for teaching with and learning from evaluative feedback," *Cognition*, vol. 167, pp. 91–106, 2017.

[23] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[24] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[25] S. Huang, R. F. J. Dossa, C. Ye, J. Braga, D. Chakraborty, K. Mehta, and J. G. Araújo, "Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms," *Journal of Machine Learning Research*, vol. 23, no. 274, pp. 1–18, 2022. [Online]. Available: http://jmlr.org/papers/v23/21-1342.html

[26] B. Eysenbach and S. Levine, "Maximum entropy rl (provably) solves some robust rl problems," *arXiv preprint arXiv:2103.06257*, 2021.