ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Hybrid-AIRL: Enhancing Inverse Reinforcement Learning with Supervised Expert Guidance

Bram Silue[1] and Santiago Amaya-Corredor[1] and Patrick Mannion[2] and Lander Willem[3] and Pieter Libin[1]

1- Vrije Universiteit Brussel - Artificial Intelligence Lab
Pleinlaan 2, Brussels - Belgium

2- University of Galway - College of Science and Engineering
University Rd, Galway - Ireland

3- University of Antwerp - Family Medicine and Population Health
Prinsstraat 13, Antwerp - Belgium

Abstract. Adversarial Inverse Reinforcement Learning (AIRL) addresses sparse rewards by inferring dense reward functions from expert demonstrations, but its performance in complex, imperfect-information settings is underexplored. We evaluate AIRL in Heads-Up Limit Hold'em (HULHE) poker and observe that it faces challenges producing sufficiently informative rewards. To address this, we introduce Hybrid-AIRL (H-AIRL), which improves reward inference and policy learning using a partially supervised loss from expert data and stochastic regularization. Experiments on Gymnasium benchmarks and HULHE poker show that H-AIRL improves sample efficiency and training stability, highlighting the value of supervised signals in inverse RL.

1 Introduction

Deep reinforcement learning (RL) has achieved strong results in many domains [1, 2], yet reward design remains a bottleneck. Sparse or delayed rewards make it difficult for RL agents to learn effective strategies in complex decision-making tasks. Poker illustrates these challenges: it is an imperfect-information, stochastic game where feedback is only observed at the end of each hand, causing standard RL methods to struggle.

Inverse Reinforcement Learning (IRL) offers an alternative by inferring reward functions from expert demonstrations [3]. Maximum-entropy IRL [4] and adversarial methods such as GAIL [5] and later AIRL [6] model expert behavior through a discriminator-policy game, with AIRL additionally recovering an explicit reward function.

Despite these capabilities, our experiments show that AIRL faces difficulties in Heads-Up Limit Hold'em (HULHE) poker, where large state-action spaces and partial observability hinder reward function inference. To address this, we propose Hybrid Adversarial Inverse Reinforcement Learning (H-AIRL), which augments AIRL with supervised expert guidance and stochastic regularization. Across Gymnasium benchmarks and HULHE poker, H-AIRL improves stability, sample efficiency, and the fidelity of inferred rewards.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

2 Background

Reinforcement Learning (RL) models control problems through a Markov Decision Process, defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where an agent seeks a policy $\pi(\mathbf{a}|\mathbf{s})$ that maximizes the expected discounted return [7]. In many domains, however, specifying the reward function $r(\mathbf{s}, \mathbf{a})$ is difficult or impractical.

IRL inverts the RL problem by inferring a reward function $f(\mathbf{s}, \mathbf{a})$ that explains expert trajectories generated by an (approximately) optimal policy [3]. Maximum-entropy IRL models expert behavior by assigning higher probability to high-reward trajectories, but computing this distribution is intractable in large state spaces, motivating adversarial alternatives.

Adversarial IRL methods treat learning from demonstrations as a game between a policy and a discriminator. While GAIL [5] imitates expert behavior without recovering a reward, AIRL [6] extends this framework by learning a reward function jointly with the policy through a discriminator that distinguishes expert from policy-generated state-action pairs. This enables AIRL to recover a reward signal that explains expert behavior.

3 The Hybrid-AIRL Framework

H-AIRL augments AIRL with supervised alignment and stochastic regularization, aiming to stabilize training and improve reward inference while remaining compatible with the adversarial IRL framework.

3.1 Policy Objective

In AIRL, the policy π_ϕ is optimized using the entropy-regularized loss

$$\mathcal{L}_{\text{AIRL}}^{\text{policy}} = -f_\theta(\mathbf{s}, \mathbf{a}) + \log \pi_\phi(\mathbf{a}|\mathbf{s}). \tag{1}$$

which encourages high-reward, high-entropy behavior under the learned reward f_θ. H-AIRL extends this by also aligning the policy's action distribution with expert demonstrations. Given expert trajectories ρ_E, we define the hybrid loss

$$\mathcal{L}_{\text{H-AIRL}}^{\text{policy}} = (1-\alpha)\,\mathcal{L}_{\text{AIRL}}^{\text{policy}} + \alpha\,\mathcal{L}_{\text{S}}, \tag{2}$$

where $\mathcal{L}_{\text{S}}$ is a supervised loss such as the cross-entropy between $\pi_\phi(\cdot|\mathbf{s})$ and the expert's empirical action distribution or the mean-squared error for continuous action vectors. The hyperparameter $\alpha \in [0, 1]$ controls the balance between adversarial IRL and direct supervision. This hybrid formulation preserves AIRL's reward-learning mechanism while injecting expert-guided supervision that improves stability.

3.2 Discriminator Objective

AIRL trains the discriminator using the binary cross-entropy loss

$$\mathcal{L}_{\text{AIRL}}^{\text{disc}} = -\mathbb{E}_{\rho_E}[\log D_\theta(\mathbf{s}, \mathbf{a})] - \mathbb{E}_{\pi_\phi}[\log(1 - D_\theta(\mathbf{s}, \mathbf{a}))], \tag{3}$$

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

which encourages the learned reward f_θ to score expert state-action pairs higher than policy samples. When ground-truth rewards r_{env} are available, H-AIRL adds a supervised regularizer, *i.e.*, $\mathcal{L}_{\text{S}}^{\text{disc}} = \mathbb{E}_{\rho_E}[(f_\theta(\mathbf{s},\mathbf{a}) - r_{\text{env}}(\mathbf{s},\mathbf{a}))^2]$, and combines both terms as

$$\mathcal{L}_{\text{H-AIRL}}^{\text{disc}} = (1-\beta)\mathcal{L}_{\text{AIRL}}^{\text{disc}} + \beta\,\mathcal{L}_{\text{S}}^{\text{disc}}, \tag{4}$$

where $\beta \in [0,1]$ controls the contribution of supervised guidance. When no ground-truth rewards are available, we set $\beta = 0$.

3.3 Stochastic Regularization

The supervised policy term can make π_ϕ mimic the expert too quickly, reducing the discriminator's feedback signal. To avoid this, we add lightweight stochastic regularization by perturbing each policy action with Gaussian noise: $\tilde{\mathbf{a}} = \mathbf{a} + \boldsymbol{\eta}$, where $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ adds noise across each action dimension, and σ follows a decaying schedule along the mini-batch dimension. This introduces a mix of near-expert and perturbed actions, preventing discriminator overfitting.

Applying this noise to both adversarial and supervised discriminator terms yields the final H-AIRL discriminator objective

$$\mathcal{L}_{\text{H-AIRL}}^{\text{disc}} = (1-\beta)\mathcal{L}_{\text{AIRL+SR}}^{\text{disc}} + \beta\,\mathcal{L}_{\text{S+SR}}^{\text{disc}}. \tag{5}$$

4 Experimental Setup

H-AIRL[1] follows the standard IRL loop: a discriminator is trained to distinguish expert and policy samples, and the policy is updated to maximize the inferred reward. After IRL training, the discriminator serves as a learned reward function, which we use to train a downstream RL agent.

4.1 Benchmarks

We evaluate AIRL and H-AIRL on five Gymnasium environments – Pendulum, Ant, HalfCheetah, LunarLander, and MountainCar [8]. The first three match the continuous-control tasks used in the original AIRL paper [6], while the latter two provide discrete-action settings. Expert demonstrations are generated with Proximal Policy Optimization (PPO) [9] or Deep Q-Networks (DQN) [1].

We further evaluate on Heads-Up Limit Hold'em (HULHE) poker, a challenging imperfect-information game with sparse terminal rewards. We use the IRC Poker dataset[2], which contains over one million HULHE state-action pairs. Because folding reveals no private cards, we cannot learn from folding actions and must train only on the remaining observable ones (call, raise, check).

[1] `https://github.com/silue-dev/hairl`

[2] `https://poker.cs.ualberta.ca/irc_poker_database.html`

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

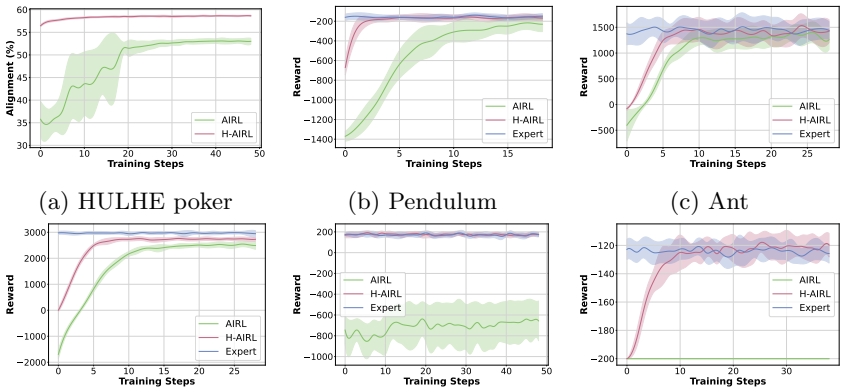
4.2 Evaluation

We assess both *policy* performance and *reward* quality. For Gymnasium tasks, we report IRL reward learning curves. For HULHE, we compute the state-wise action alignment, which is the fraction of states where the learned policy matches the expert. To evaluate reward quality, we train RL agents using only the IRL-derived reward and measure their environment-reward performance. Poker rewards are integrated through RLCard's DQN implementation [10]. All curves report mean and standard deviation across 10 random seeds.

For poker, we also run 1,000,000 tournaments across 20 seeds, comparing DQN agents trained with dense IRL rewards to those using standard sparse payoffs. Performance is measured in milli-big-blinds per hand (mbb/h).

5 Experimental Results

We first examine IRL training. Figure 1 shows the learning curves of AIRL and H-AIRL on the Gymnasium benchmarks and HULHE poker.



(a) HULHE poker (b) Pendulum (c) Ant

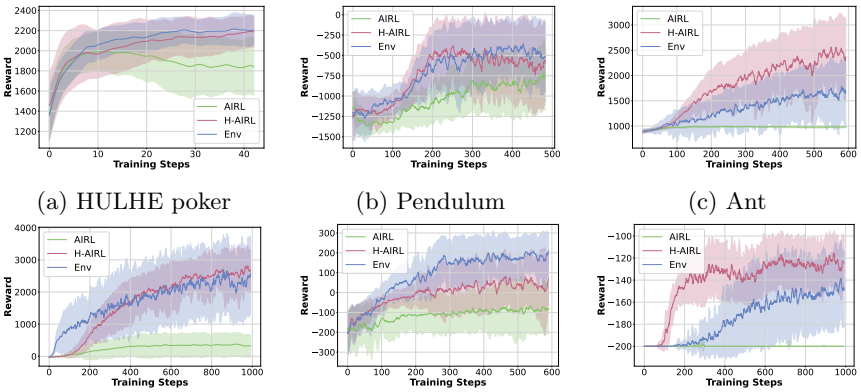
(d) HalfCheetah (e) LunarLander (f) MountainCar

Fig. 1: Learning curves for AIRL (green) and H-AIRL (red) on Gymnasium benchmarks and HULHE poker, alongside an expert PPO baseline (blue).

Across these benchmarks, H-AIRL converges more quickly and exhibits reduced variance relative to AIRL, indicating more stable and efficient reward learning.

Next, we evaluate RL agents that use the learned rewards, an important step for assessing the quality of the inferred reward function. Figure 2 shows the environment-reward learning curves of agents trained using rewards from AIRL and H-AIRL.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.



(a) HULHE poker (b) Pendulum (c) Ant

(d) HalfCheetah (e) LunarLander (f) MountainCar

Fig. 2: RL training curves of an agent using environment (blue), AIRL-derived (green), and H-AIRL-derived (red) rewards across benchmarks.

In most tasks, agents trained with H-AIRL rewards reach higher returns or converge more rapidly than those trained with AIRL rewards, and generally match or approach performance under the environment reward.

For poker, Table 1 shows results from 1,000,000 tournaments of AIRL-DQN and H-AIRL-DQN against RLCard's default DQN agent. AIRL-DQN performs significantly worse than the baseline DQN, with a large negative payoff, whereas H-AIRL-DQN achieves a positive payoff and competitive performance. For context, professional poker players consider 50 mbb/h a meaningful margin [11], underscoring the practical significance of these differences.

Model	Payoff (mbb/h)	p-value
H-AIRL-DQN	$+96 \pm 14$	$< 10^{-10}$
AIRL-DQN	-693 ± 34	$< 10^{-10}$

Table 1: Performance of AIRL-DQN and H-AIRL-DQN against DQN in HULHE poker, measured as average payoff and standard error in mbb/h.

6 Discussion

Our experiments show that H-AIRL consistently improves over AIRL in both policy learning and reward inference. Across the Gymnasium benchmarks, H-AIRL converges more quickly, exhibits greater training stability, and achieves stronger alignment with expert behavior compared to AIRL. In HULHE poker, where sparse rewards and partial observability pose significant challenges, H-AIRL recovers reward functions that better guide downstream RL agents, as reflected in improved RL learning curves and substantially higher tournament payoffs. These results indicate that while AIRL remains a strong and widely

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

used foundation, incorporating supervised structure and stochastic regularization yields a more stable and informative IRL framework in complex domains.

Our aim in this work is to isolate and evaluate the contribution of H-AIRL's hybrid-loss framework relative to the foundational AIRL baseline. A broader empirical evaluation against recent IRL methods is left to future work in order to better situate H-AIRL within the current landscape. While our results are encouraging, the study has several limitations. First, our poker data lacks folding actions, a common limitation in real-world datasets where folded cards are never revealed. Second, H-AIRL does not recover disentangled rewards that lead to theoretical guarantees for transfer, and does not explicitly address partial observability. These points suggest several directions for future work, such as extending the hybrid framework formulation to produce disentangled rewards, and studying recurrent or belief-state extensions of H-AIRL for partially observable domains.

References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.

[2] Pieter J. K. Libin, Arno Moonens, Timothy Verstraeten, Fabian Perez-Sanjines, Niel Hens, Philippe Lemey, and Ann Nowé. Deep reinforcement learning for large-scale epidemic control. In Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke, editors, *Proceedings of ECML-PKDD*, 2021.

[3] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In Pat Langley, editor, *Proceedings of ICML*, 2000.

[4] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of AAAI*, 2008.

[5] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Proceedings of NeurIPS*, 2016.

[6] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *Proceedings of ICLR*, 2018.

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

[8] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, 2017.

[10] Daochen Zha, Kwei-Herng Lai, Songyi Huang, Yuanpu Cao, Keerthana Reddy, Juan Vargas, Alex Nguyen, Ruzhe Wei, Junyu Guo, and Xia Hu. Rlcard: A platform for reinforcement learning in card games. In *Proceedings of IJCAI*, 2020.

[11] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 2017.