Pertanika J. Sci. & Technol. 34 (1): 457 - 470 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Short Communication

Beyond Boundaries: Exploring the Expanding Horizons of Inverse Reinforcement Learning

Ojonukpe Sylvester Egwuche[1,2], Olanrewaju Victor Johnson[2], Arome Junior Gabriel[3], and Chew XinYing[4]*

[1]*Unit for Data Science and Computing, North-West University, 2520 Potchefstroom, South Africa*
[2]*Department of Computer Science, Federal Polytechnic, PMB 727, Ile-Oluji, Ondo State, Nigeria*
[3]*Department of Cybersecurity, Federal University of Technology, PMB 704, Akure, Nigeria*
[4]*School of Computer Sciences, 11800 Unversiti Sains Malaysia, Penang, Malaysia*

ABSTRACT

Reinforcement learning (RL) has achieved significant success in complex, sequential decision-making tasks. However, it remains constrained by its dependence on predefined reward functions, limiting adaptability in dynamic environments. Inverse Reinforcement Learning (IRL) addresses this limitation by inferring reward structures from expert demonstrations, enabling more flexible and context-aware agents. The study explores the potential of IRL's in enhancing the efficiency and adaptability of modern autonomous systems. The pivotal role of IRL in modelling human-like reasoning and imagination is examined across domains, including robotics, autonomous driving, personalised medicine, and cybersecurity, alongside discussion on current solutions, challenges, and emerging research directions. The findings underscore future improvements for human cognitive capabilities and machine autonomy.

Keywords: Agentic AI, autonomous systems, cybersecurity, robotics, inverse reinforcement learning, explainable AI, ethical AI

ARTICLE INFO

Article history:
Received: 15 April 2025
Accepted: 22 August 2025
Published: 25 February 2026

DOI: https://doi.org/10.47836/pjst.34.1.20

E-mail addresses:
ojoegwuche@fedpolel.edu.ng (Ojonukpe Sylvester Egwuche)
olajohnson@fedpolel.edu.ng (Olanrewaju Victor Johnson)
ajgabriel@futa.edu.ng (Arome Junior Gabriel)
xinying@usm.my (Chew XinYing)
* Corresponding author

INTRODUCTION

Beyond the traditional reinforcement learning, how can artificial intelligence systems learn human preferences without explicit definition of the reward function? The traditional Reinforcement Learning (RL) is a forward learning approach that searches for the optimal policy that returns

e-ISSN: 2231-8526

maximal reward value in limited action spaces (Egwuche et al., 2025). The learning approach requires the agent to interact with the environment exhausting all avenues to learn and receive feedback in the form of rewards or punishments (Jayaraman et al., 2024). In complex environments such as autonomous driving and robotics where the reward functions are not explicitly stated, RL agents are observed to find it difficult making computationally efficient decisions that return maximum reward functions (Zhang et al., 2022). Hence, Inverse Reinforcement Learning (IRL) has emerged by relying on traditional RL to change the thinking of Artificial Intelligence (AI) without relying on explicitly defined reward function in learning human preferences. The IRL deduces reward functions from observed behaviours rather than handcrafting them. According to Shah and De Pietro in (Shah & De Pietro, 2021), IRL is modelled as a Markov Decision Process (MDP) that enables autonomous agents to simulate actions in Intelligent Environments (IEs). Consequently, it allows systems to self-improve from data without explicit programming, thereby, establishing it as a robust technique for learning from demonstrations (Das et al., 2021). For instance, a child observing an adult climbing the stairs might begin with her right leg, earning a reward (possibly clapping) in traditional RL for each step or when she reaches the top, as shown in Figure 1. However, IRL relies on successful techniques to deduce a flexible reward function rather than strict imitating if she adjusts by using her left leg. This adaptability enables IRL's appeal in constructing agents that learn from recorded task performances in IEs without interference, broadening RL's scope to tasks where reward specification (e.g., +1 for success, -1 for failure) is impractical or conflicted (Arora & Doshi, 2021).

Inverse RL is fundamental in building human imagination in AI-powered decision support systems. (Liu et al., 2024). Furthermore, IRL shines in its ability to share an agent's reward preferences succinctly with another in identical settings. It was discovered that early successes in traffic trajectory optimisation have evolved into recent advances like human gait analysis, driving styles, and expert video game play with IRL. However, the assumed environmental and goal alignment in the examples could render transferred rewards useless, underscoring a key limitation in IRL's practical reach if care is not taken. In contrast to RL's focuss on learning optimal behaviour from reward-driven experiences, IRL flips the script, using observed state-action trajectories to uncover the underlying reward function. While linear programming suits simpler MDP problems, dynamic programming tackles complex cases, though both grapple with the ill-posed challenge of multiple reward functions explaining the same data. With MDP remaining the predominant model for IRL, variants such as Hidden-Parameter MDPs (HPMDP) and Partially Observable MDPs (POMDP) have emerged to handle more complex scenarios (Arora & Doshi, 2021; Korsunsky et al., 2019). Moreover, ML approaches are game changers for IRL applications. However, how successful were these approaches lies in mitigating the challenges attributed to IRL?

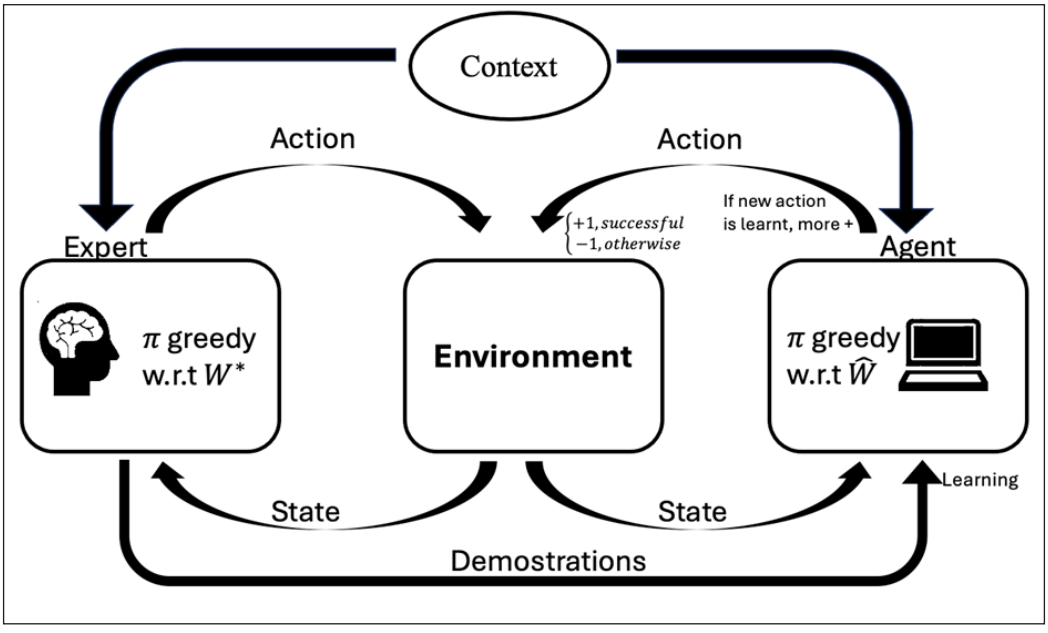

Figure 1. IRL Workflow showing example of Contextual MDPs
(Adapted: Belogolovsky et al., 2021)

While exploring the IRL model in subsequent section, including fundamental bottlenecks in its design and applications, a succinct focusses on key emerging areas in IRL adoption that could further spur research interest were addressed. Reinforcement learning and IRL are dual problems operating within the MDP framework. While RL aims at deriving the policy given the rewards, IRL aims at learning the rewards given the policy. Reverse-engineering the reward function forms the foundation of IRL.

IRL SYSTEMS

Markov Decision Process and RL are foundational in the development of IRL. Simply state, the MDP is a 5-tuple framework – $M = (S, A, T, R, \gamma)$, where S denotes a finite set of situational states of the agent, A, a finite set of actions taken by agent, T representing the transitions $T(s^{'}|s, a)$ defined as $T: S \times A \rightarrow Prob(S)$, R denotes a *reward function* expressed as $R(s, a, s^{'})$, and γ as discounting factor determining the future rewards expressed as $\gamma \in [0,1]$, both policy and value function are derived. A policy function π maps states to actions, dictating the agent's behaviour as deterministic: $\pi(s) = a$ or stochastic: $\pi(s)$, whereas the value function $V_{\pi}(s)$ denoting the expected long-term cumulative R, an agent receives when starting from state s and following policy π expressed as:

$$V_\pi(s) = E_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s\right] \quad [1]$$

In reinforcement learning, the reward R is known but the optimal policy $\pi*$ is unknown.

$$\pi^* = \arg\max_\pi E_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right] \quad [2]$$

However, in IRL, the reward R is unknown and the expert's demonstrations are observed to derive the reward function such that the expert's policy π_E is approximately optimal expressed as:

$$\pi_E = arg\max_\pi E_\pi\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right] \quad [3]$$

Given $\phi(s, a)$ as a feature vector, the expert's feature expectations under policy π is given as:

$$\mu_E = E_{\pi_E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t)\right] \quad [4]$$

In various application domains, the practical implications of the equations can be clearly illustrated. For instance, in an autonomous driving scenario, states s can represent the positions of the vehicle, the road conditions or the traffic signals while actions a can denote wait, accelerate, brake or turn. The features $\phi(s, a)$ denote the distance from other vehicles, lane positions etc. and the expert policy π_E captures the demonstrations of human drivers. The goal of IRL here is to learn policies that mimic human drivers such as preference of safety over speed. Similarly, in network security, states s can represent system logs, network traffic trend or intrusion alerts while actions a can denote block, allow or escalate etc. The features $\phi(s, a)$ are derived from indicators such as traffic volume, packet drop rates or the number of failed logins, and the expert Behaviour represents the response strategies of human security analyst. The objective of IRL in this scenario is to extract expert strategies for automatic decision-making in threat mitigation. These examples demonstrate how IRL-derived value functions bridge mathematical representations with observable real-world performance.

However, the practicality of IRL is immediately confronted by challenges highlighted in Ng and Russell's critique, as discussed by Arora and Doshi (2021). The authors emphasise that even trivial functions (e.g., all zeros) can serve as valid fits, revealing the inherent

difficulty of accurate reward inference. While evaluation metrics such as the closeness of the learned reward function, behaviour accuracy, and Inverse Learning Error (ILE) offer partial remedies, no single measure fully resolves this limitation. Beyond accuracy, IRL also struggles with generalisation, as expert demonstrations typically cover only a subset of states. Extrapolating to unseen scenarios introduces the risk of significant errors, especially under sparse data conditions.

Reward functions are often represented as weighted combinations of features, simplifying the problem to one of parameter tuning to address this consequent. Yet, the fidelity of such an approach critically depends on selecting features that faithfully capture the expert's intent. Critics argue that this reliance on prior knowledge, compounded by the limited coverage of demonstrations, reduces IRL's robustness compared to supervised and unsupervised learning approaches, which enjoy broader applicability. Aside, computational complexity of IRL was presented as an additional barrier to IRL widespread adoption. As IRL relies on MDP, the iterative solution to the latter often becomes computationally intractable as the problem size increases, giving rise to the curse of dimensionality, particularly in continuous state-action spaces such as robotics (Ren et al., 2024).

To overcome these challenges, studies show that the efficiency of IRL could be enhanced by applying function approximation techniques such as deep Q-learning (Xue et al., 2021), gradient-based optimisation, trajectory-based learning, transfer learning, and meta-learning, OpenAI Gym (Yang et al., 2024). Other applicable techniques in literature include Distributed-IRL frameworks, such as RLlib and DeepMind's SEED, are also utilised to improve the efficiency of IRL. These methods support large-scale parallelisation. In attaining different levels of trade-offs in terms of efficiency and sample complexity, specialised methods have also been developed. These methods include Deep-IRL (D-IRL), Hierarchical-IRL (H-IRL), Maximum-Entropy-IRL (ME-IRL), Bayesian-IRL (BIRL), Monte Carlo Tree Search (MCTS), Policy-Gradient-IRL, Trust Region Policy Optimisation (TRPO), and Adversarial-IRL (A-IRL). Recent studies show that these techniques are helpful to innovating AI solutions aimed at addressing specific computational and application-driven challenges. For instance, Cooperative-IRL (C-IRL), allows multiple agents work together by adapting learned reward functions to share goals. Whereas Multi-Agent IRL (MA-IRL) and Multi-Agent Adversarial IRL (MAA-IRL) extend IRL into collaborative and competitive environments, which are applicable in modern robotics and autonomous systems. Life Long-IRL (LL-IRL) focusses on continual learning, which enable agents to refine their learned reward functions over time as new data becomes available, Model-Free IRL (MF-IRL) seeks to bypass explicit environment representation. The former reduces reliance on static demonstration, while the latter enhances scalability in high-dimensional spaces. The aspect of Maximum Causal Entropy IRL (MCE-IRL) provides a probabilistic framework that accounts for the uncertainty and stochasticity inherent in

expert behaviour, making it a more robust alternative to conventional ME-IRL (Ren et al., 2024). The emergence of modified variants of IRL for high-level efficiency has inspired the rapid applications of the learning paradigm in different critical areas.

GROWING INTERESTS IN IRL

A wide range of applications is becoming practicable with IRL. While humanoid robots continue to see competitive advancements in design and functionality, the underlying principle of behavioural analysis, learning through demonstrations, extends far beyond robotics. Personalised medicine (PM) is revolutionising healthcare by shifting the focuss towards individualised care, public engagement, and sustainable economic models. At its core, PM integrates RL, while present techniques focuss on IRL to refine treatment strategies, diagnostics, and preventive measures. RL significance reflects in dynamic treatment regimens (DTRs) for chronic disease and critical care (Shah et al., 2023). However, IRL adoption may infer optimal treatment policies from expert demonstrations, bypassing the limitations of predefined reward functions. This method works well in intricate clinical situations where complicated decision-making is required due to high-dimensional data and changing patient circumstances. While MA-IRL is useful for AI-to-human healthcare teams, challenges such as sample complexity, ethical concerns, and the opacity of deep RL models remain significant barriers to its widespread adoption (Arora & Doshi, 2021). Improving Machine Learning (ML) with Q-learning methods is a positive direction for deep RL issues in PM.

Another area of healthcare where RL-based models surpass conventional ML methods is the fields of radiology and dermatology. Moreover, IRL is used in these fields to imitate human diagnostic reasoning rather than rigid adherence to predefined reward metrics. For example, MCE-IRL captures the stochastic nature of clinical decision-making to obtain better accurate, and robust predictions (Benac et al., 2024). Nevertheless, ongoing area of research includes Federated IRL to enable decentralising learning from expert data while safeguarding patient privacy (a critical consideration in medical AI), interpretability to translate insights into actions, and generalisable strategies across diverse patient populations, ensuring that AI-driven diagnostics are both reliable and transparent. Aside, patient-specific applications, IRL broadens healthcare processes, including drug discovery, clinical trial design, and resource allocation by focussing on optimising policies without predefined reward structures, drug formulation, and dosage determination. The main challenges are scalability and regulatory hurdles (Yu et al., 2021). However, the integration of IRL with causal inference techniques could enhance policy robustness in real-world scenarios.

Looking at Natural Language Processing (NLP), Large Language Models (LLMs) have been transformational in recent time. Though M-ARL addresses coordination challenges in Multi-Agent Systems (MAS), including decentralised training, enhancing

multi-turn dialogue generation, and real-time adaptability (Sun et al., 2024), IRL refines LLMs by inferring human-like reward structures. IRL further improves context retention, conversational coherence, and ethical decision-making, which are lacking in traditional RL (Li et al., 2019). Additionally, IRL-driven open-domain dialogue systems, such as chatbots and automated help desks, learn latent reward functions from expert dialogues, allowing them to emulate nuanced human interactions without rigid objectives. For example, DeepSeek-R1 model, employing Group Relative Policy Optimisation (GRPO), enhances dialogue coherence and relevance, though challenges like data dependency and bias inheritance persist (Liu et al., 2024). Parmar and Govindarajulu (2025) proposed Supervised Fine-Tuning (SFT) in the RL pipeline to mitigate these issues, however, IRL's potential for improvement cannot be overemphasised.

Future studies may integrate scalable MARL-IRL hybrid models to optimise agent-based LLMs to address issues such as non-stationarity, credit assignment, and policy convergence at Figure 2. Other aspects of IRL focuss in NLP include text summarisation (Jian, 2022). For instance, ME-IRL may enhance sentiment-adaptive responses, making chatbot interactions more engaging and reducing generic or misleading replies. H-IRL and A-IRL advance conversational AI toward goal-oriented and emotion-aware interactions, improving user experience (Yang, 2020). In opinion mining, IRL infers implicit user preferences from expert-annotated sentiment data, enhancing adaptability across domains, dialects, and evolving linguistic patterns. Similarly, IRL improves semantic preservation in summarisation and translation by learning from high-quality, human-generated texts rather than optimising token-based objectives. However, computational costs, sample inefficiency, and real-world robustness remain key hurdles, which were earlier discussed in this paper.

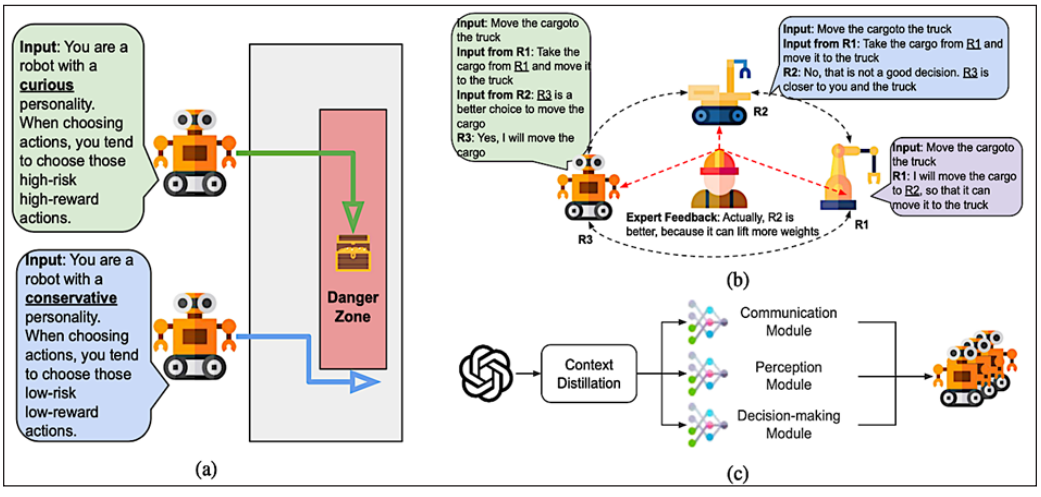

Figure 2. MARL-IRL application in LLM (Sun et al., 2024)

In optimising the functionality of autonomous vehicles, IRL plays a critical role. In the traditional RL, the agent incurs risky and unsafe behaviours such as speeding or running against the traffic, in addition to taking a longer time to figure out how to respond to yellow lights, pedestrian crossings, or sharp turns. Invariably, the agent requires a long loop of training to learn good driving policies. This is computationally intensive and not resource efficient. With IRL, the agents derive some level of understanding of the expert intentions for taking certain decisions. For example, a human driver may reduce speed when approaching a yellow light, a sharp turn, or a pedestrian crossing to avoid breaking traffic rules or cause and form of accident. This approach provides a contextual adaptation that allows the agent to adjust its driving policy in different traffic situations more naturally and intuitively (Jara-Ettinger, 2019).

In the optimisation process, automatic driving decisions are made in a manner that is safe, efficient, and similar to how a human driver would naturally behave on the road, a success attributed to predictive behaviour planning in IRL (Huang et al., 2023). A similar tendency intriguing outcome of traditional RL in robotics and human-robot-humanoid, including interaction, exploration overhead, and human-machine context, yet long training time and context-aware human-machine interaction remain challenging (Gharbi et al., 2024). In complex tasks such as arranging books in a shelf or folding laundry, the robot, with its lack of natural understanding of human preferences, may experiment with different actions to identify the policy that returns the expected rewards. Reverse engineering in robotic agents with IRL are positive direction to overcome the long training loop of trial-and-error required for the traditional RL.

Another area of interest is cognitive neuroscience, where defining explicit rewards for mental processes remains an open challenge. Researchers are constantly seeking better solutions to uncertainty in brain-related problems. IRL-based models are continuously helping to decode the neural representations of decision-making by analysing fMRI and EEG data. The study in (O'Doherty et al., 2003) uses IRL to uncover the reward structures that guide human choices in conditions such as Parkinson's disease and schizophrenia, where reward-based learning is impaired. Similarly, Redish et al. (Redish et al., 2008) applied IRL models to investigate habit formation, addictions, and impulse control disorders. We observed that the approach offers a quantitative understanding of ill-adaptive decision-making patterns with the ability to infer optimal behaviour from observations, making it an invaluable tool for brain-computer interactions (BCIs).

For instance, for patients with motor impairments that require personalised assistance, IRL-driven BCIs is helpful in observing from the demonstration of human experts to tailor neural control strategies for prosthetic limbs and communication devices, an adaptability crucial for neurorehabilitation. Furthermore, the potential of sophisticated AI technologies such as IRL for personalised neuroscience is instrumental to the breakthroughs in precision

medicine (Onciul et al., 2025). Traditional neuro-imaging techniques fundamentally rely on a supervised learning approach with predefined labels. However, IRL follows a more flexible pattern by learning from unstructured data, making it possible to deduce significant hidden cognitive states. The arguments in Marwood et al. (2018) indicate that IRL helps in understanding the neural correlates of moral decision-making and social-cognition, thereby providing insights into the brain processes related to ethical dilemmas and interpersonal interactions. Understanding the neural pathways could facilitate the recovery processes in providing target treatments associated with post-traumatic stress disorder (PTSD) and depression mapping, underscoring the significant of IRL in transforming neuroscience.

The last but not the least interesting discussion in this paper is cyber threats (CTs). As CTs continue to grow in complexity, static rule-based security models struggle to keep up with the latest adversaries. In essence, IRL promises an alternative by allowing defensive mechanisms to learn from attack patterns and anticipate the best possible solutions. How? The cyber adversary system optimises its strategies to maximise hidden rewards, whether for data exfiltration, system disruption, or privilege escalation. Then, IRL is used to provide a way to uncover the underlying strategy by analysing historical attack data. With this kind of dynamics, security frameworks can proactively counter an attack rather than react after the breach of security. In recent time, IRL has been instrumental in modern intrusion detection systems (IDS), including automating malware analysis and network traffic monitoring, misleading cyber attackers by dynamically altering system behaviour to increase uncertainty in their reward functions, and effectively reducing the success rate of adversarial exploits (Parras et al., 2022). In another dimension, IRL could enhance AI robustness by learning from adversarial interactions and adjusting defensive strategies accordingly, a case whereby classifiers resist adversarial perturbations and maintain reliability in high-risk environments (Mahjoub et al., 2024).

Researchers are also leveraging IRL to develop anomaly detection systems capable of distinguishing sophisticated, previously unseen attack patterns, which are crucial in financial fraud detection (Zelman et al., 2024). IRL-driven frameworks are ongoing in areas such as real-time risk assessments and cloud security dynamics. While the adoption of IRL in human-in-the-loop systems is on the rise, some challenges that limit the efficiency of the technique especially in highly complex and dynamic environments has been identified.

CHALLENGES AND LIMITATIONS

While the integration of IRL into human-in-the-loop systems such as autonomous driving assistance has been largely successful, other challenges that contend with the growing adoption of the learning paradigm has been noted. These challenges stem from the heavy reliance on expert trajectories for reward inference, especially in environments where collecting the optimal trajectories that reflect the intended objectives are infeasible.

The reliance on expert trajectories can lead to security breaches where attackers can mislead policy training via data poisoning. For instance, in critical domains such as in the healthcare sector and autonomous driving vehicles, a mis-specified reward function can cause direct harm. Ensuring fairness, safety, and compliance with ethical guidelines requires explainability and ongoing monitoring. Without transparent reasoning and adequate verification of decision-making process, IRL solutions may not be embraced in sensitive domains.

Another challenge that has limited the effectiveness of IRL is reward ambiguity. Reward ambiguity is experienced when multiple reward functions are associated with the same expert demonstration, making it difficult to know the exact reward function that corresponds to the intention of the expert. For example, in NLP, if demonstrations (e.g., human dialogue transcripts) come from sources that contain ambiguous stereotypes, an IRL-trained chatbot may infer a reward that rewards responses that align with those biases. Other challenges include dependency on environmental models, computational and sample inefficiencies in demonstrations-scarce environments or noisy demonstrations and assumption of optimal expert demonstrations. To overcome the ambiguity, extra constraints, such as prior knowledge or preference learning that avoid unintended outcomes could be incorporated in the design processes. Addressing these challenges to accommodate contradictory human data opens new research directions for further investigation in strengthening the efficiency of IRL solutions.

CONCLUSION

The paper discusses the flexibility and transformative power of IRL in several applications and realms such as self-driving, robotics, PM, NLP, cognitive neuroscience, and cybersecurity. It is observed that IRL techniques allow agents to easily predict reward functions based on expert demonstrations and do not have to depend on manually designed reward functions. The key takeaways are that IRL offers benefits such as adaptive decision making, cross-domain applicability, integration with modern AI techniques, and even cyber or information security relevance. On directions of further research, there are several open research questions. What is needed for scaling IRL to multi-modal demonstrations, enhancing video with audio and textual information? How are sample complexities to be reduced without sacrificing policy accuracy? What should some of the main ways to incorporate Explainable AI principles into IRL frameworks be so that ethical and transparent decision-making can be made? How do you think IRL will best be used in safety-critical areas, especially autonomous systems and cybersecurity, with stringent regulation and operational restrictions? Answering these questions will be essential to make the best of IRL in academic and industrial use.

ACKNOWLEDGEMENT

This work is funded by Universiti Sains Malaysia, Short Term Grant [Grant Number: 304/PKOMP/6315616], for the Project entitled "New Coefficient of Variation Control Charts based on Variable Charting Statistics in Industry 4.0 for the Quality Smart Manufacturing and Services" and Unit Data for Science and Computing, North-West University, South Africa.

ABBREVIATIONS

Term	:	Definition
RL	:	Reinforcement Learning
IRL	:	Inverse Reinforcement Learning
AI	:	Artificial Intelligence
MDP	:	Markov Decision Process
IEs	:	Intelligent Environments
POMDP	:	Partially Observable MDPs
HPMDP	:	Hidden-Parameter MDPs
ILE	:	Inverse Learning Error
D-IRL	:	Deep-IRL
H-IRL	:	Hierarchical-IRL
MCTS	:	Monte Carlo Tree Search
BIRL	:	Bayesian-IRL
ME-IRL	:	Maximum-Entropy-IRL
TRPO	:	Trust Region Policy Optimisation
A-IRL	:	Adversarial-IRL
C-IRL	:	Cooperative-IRL
MA-IRL	:	Multi-Agent IRL
MAA-IRL	:	Multi-Agent Adversarial IRL
MF-IRL	:	Model-Free IRL
LL-IRL	:	Life Long-IRL
MCE-IRL	:	Maximum Causal Entropy IRL
NLP	:	Natural Language Processing
LLMs	:	Large Language Models
GRPO	:	Group Relative Policy Optimisation
SFT	:	Supervised Fine-Tuning

REFERENCES

Arora, S., & Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods, and progress. *Artificial Intelligence, 297*, Article 103500. https://doi.org/10.1016/j.artint.2021.103500

Belogolovsky, S., Korsunsky, P., Mannor, S., Tessler, C., & Zahavy, T. (2021). Inverse reinforcement learning in contextual MDPs. *Machine Learning, 110*(9), 2295-2334. https://doi.org/10.1007/s10994-021-05984-x

Benac, L., Sharma, A., Parbhoo, S., & Doshi-Velez, F. (2024). Inverse transition learning: Learning dynamics from demonstrations [Preprint]. *arXiv*.

Das, N., Bechtle, S., Davchev, T., Jayaraman, D., Rai, A., & Meier, F. (2021). Model-based inverse reinforcement learning from visual demonstrations. In *Proceedings of the 5th Conference on Robot Learning (CoRL 2021)*. PMLR.

Egwuche, O. S., Greeff, J., & Ezugwu, A. E. (2025). Optimised task offloading in multi-domain IoT networks using distributed deep reinforcement learning in edge computing environments. *IEEE Access*. Advance online publication. https://doi.org/10.1109/ACCESS.2025.3535870

Gharbi, H., Elaachak, L., & Fennan, A. (2024). Replicating video game players' behaviour through deep reinforcement learning algorithms. *Journal of Theoretical and Applied Information Technology, 102*(15), 5734-5749.

Huang, Z., Liu, H., Wu, J., & Lv, C. (2023). Conditional predictive behaviour planning with inverse reinforcement learning for human-like autonomous driving. *IEEE Transactions on Intelligent Transportation Systems, 24*(7), 7244-7258. https://doi.org/10.1109/TITS.2023.3254579

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioural Sciences, 29*, 105-110. https://doi.org/10.1016/j.cobeha.2019.04.010

Jayaraman, P., Desman, J., Sabounchi, M., Nadkarni, G. N., & Sakhuja, A. (2024). A primer on reinforcement learning in medicine for clinicians. *NPJ Digital Medicine, 7*(1), Article 337. https://doi.org/10.1038/s41746-024-01316-0

Jian, H. Z. (2022). Text summarisation for news articles by machine learning techniques. *Applied Mathematics and Computational Intelligence (AMCI), 11*(1), 174-196.

Korsunsky, P., Belogolovsky, S., Zahavy, T., Tessler, C., & Mannor, S. (2019). Inverse reinforcement learning in contextual MDPs [Preprint]. *arXiv*.

Li, Z., Kiseleva, J., & De Rijke, M. (2019). Dialogue generation: From imitation learning to inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/aaai.v33i01.33016722

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., & Ruan, C. (2024). DeepSeek-V3 technical report [Preprint]. *arXiv*

Liu, Y., Zhang, R., Du, H., Niyato, D., Kang, J., Xiong, Z., & Kim, D. I. (2024). Defining problem from solutions: Inverse reinforcement learning (IRL) and its applications for next-generation networking [Preprint]. *arXiv*.

Mahjoub, C., Hamdi, M., Alkanhel, R. I., Mohamed, S., & Ejbali, R. (2024). An adversarial environment reinforcement learning-driven intrusion detection algorithm for Internet of Things. *EURASIP Journal*

on Wireless Communications and Networking, 2024(1), Article 21. https://doi.org/10.1186/s13638-024-02348-6

Marwood, L., Wise, T., Perkins, A. M., & Cleare, A. J. (2018). Meta-analyses of the neural mechanisms and predictors of response to psychotherapy in depression and anxiety. *Neuroscience & Biobehavioural Reviews, 95*, 61-72. https://doi.org/10.1016/j.neubiorev.2018.09.022

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 38*(2), 329-337. https://doi.org/10.1016/S0896-6273(03)00169-7

Onciul, R., Tataru, C.-I., Dumitru, A. V., Crivoi, C., Serban, M., Covache-Busuioc, R.-A., Radoi, M. P., & Toader, C. (2025). Artificial intelligence and neuroscience: Transformative synergies in brain research and clinical applications. *Journal of Clinical Medicine, 14*(2), Article 550. https://doi.org/10.3390/jcm14020550

Parmar, M., & Govindarajulu, Y. (2025). Challenges in ensuring AI safety in DeepSeek-R1 models: The shortcomings of reinforcement learning strategies [Preprint]. *arXiv*.

Parras, J., Almodóvar, A., Apellániz, P. A., & Zazo, S. (2022). Inverse reinforcement learning: A new framework to mitigate an intelligent backoff attack. *IEEE Internet of Things Journal, 9*(24), 24790-24799. https://doi.org/10.1109/JIOT.2022.3194694

Redish, A. D., Jensen, S., & Johnson, A. (2008). Addiction as vulnerabilities in the decision process. *Behavioural and Brain Sciences, 31*(4), 461-487. https://doi.org/10.1017/S0140525X08004986

Ren, J., Swamy, G., Wu, Z. S., Bagnell, J. A., & Choudhury, S. (2024). Hybrid inverse reinforcement learning [Preprint]. *arXiv*.

Shah, S. I. H., & De Pietro, G. (2021). An overview of inverse reinforcement learning techniques. In *Proceedings of Intelligent Environments 2021* (pp. 202-212). IOS Press. https://doi.org/10.1007/s10489-022-04173-0

Shah, S. I. H., De Pietro, G., Paragliola, G., & Coronato, A. (2023). Projection-based inverse reinforcement learning for the analysis of dynamic treatment regimes. *Applied Intelligence, 53*(11), 14072-14084.https://doi.org/10.1007/s10489-022-04173-0

Sun, C., Huang, S., & Pompili, D. (2024). LLM-based multi-agent reinforcement learning: Current and future directions [Preprint]. *arXiv*.

Xue, W., Lian, B., Fan, J., Kolaric, P., Chai, T., & Lewis, F. L. (2023). Inverse reinforcement Q-learning through expert imitation for discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems, 34*(5), 2386-2399. https://doi.org/10.1109/TNNLS.2021.3106635

Yang, B., Lu, Y., Wan, R., Hu, H., Yang, C., & Ni, R. (2024). Meta-IRLSOT++: A meta-inverse reinforcement learning method for fast adaptation of trajectory prediction networks. *Expert Systems with Applications, 240*, Article 122499. https://doi.org/10.1016/j.eswa.2023.122499

Yang, Z. (2020). Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)* (pp. 193-202). IEEE. https://doi.org/10.1109/CVPR42600.2020.00027

Yu, C., Liu, J., Nemati, S., & Yin, G. (2021). Reinforcement learning in healthcare: A survey. *ACM Computing Surveys, 55*(1), Article 1. https://doi.org/10.1145/3477600

Zelman, J., Stefanik, M., Weiss, M., & Teichmann, J. (2024). Adversarial inverse reinforcement learning for market making. In *Proceedings of the 5th ACM International Conference on AI in Finance* (pp. 81-89). Association for Computing Machinery. https://doi.org/10.1145/3677052.3698641

Zhang, R., Xiong, K., Tian, X., Lu, Y., Fan, P., & Letaief, K. B. (2022). Inverse reinforcement learning meets power allocation in multi-user cellular networks. In *Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 1-2). IEEE. https://doi.org/10.1109/INFOCOMWKSHPS54753.2022.9798257