International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 09 | Sept - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Bayesian CIRL: A Unified Framework for Adaptive and Trustworthy Human-Agent Collaboration

Dr C Bhuvaneshwari[1], Dr R Gayathri[2]

[1]Department of Computer Science & Dr G R Damodran College of Science(Autonomous),Coimbatore, India
[2]Department of Computer Science(PG) & Kristu Jayanti (Deemed to be University), Bengaluru, India

---

Abstract - The rapid development of agentic AI, where autonomous agents act instead of humans need to align with human values to attain their objectives, despite initially having limited knowledge of those values.Using traditional reinforcement learning in agentic AI may result in incorrectly formulated reward functions resulting in unsafe behaviours. Furthermore RL agents encounter difficulties in sampling, reward misalignment, and lack of cross-domain generalization. In this paper, we introduce Bayesian Cooperative Inverse Reinforcement Learning (Bayesian CIRL). This is a new method that uses a latent variable with a probabilistic belief about the distribution of the human reward function. Unlike the traditional IRL, Bayesian CIRL treats value alignment as a cooperative game, where the agent updates its beliefs by observing human actions and also plans its own actions to maximize the uncertain joint reward. This approach allows adaptive cooperation through active learning. It also transfer the learned behaviour to reduce uncertainty over human inclinations. Experimental tests show that Bayesian CIRL offers more robust and accurate value alignment than standard IRL algorithms. It handles ambiguity better and allows for reliable interaction between humans and agents. This framework provides a clear way to introduce agentic AI systems that align with human ethics and societal expectations.

Key Words: Agentic AI, Reinforcement Learning, Inverse Reinforcement Learning, Cooperative Inverse Reinforcement Learning, Bayesian CIRL

1. INTRODUCTION *(Size 11, Times New roman)*

Recent developments in advanced technologies have led to a deeper integration of intelligent agents into everyday life and social structures. This integration has created new multi-agent systems where humans and autonomous agents interact and collaborate in unprecedented ways. Such interactions reshape our understanding of social dynamics and challenge traditional concepts of agency and decision-making in complex environments (Ajmeri et al., 2020). Values are fundamental forces that influence attitudes and guide human behaviour (Gabriel, 2020). These values are not inherent but are shaped by cultural and societal influences, as well as significant life experiences. Behavior is both shaped by and influences values in a reciprocal manner. The importance of specific values in guiding behavior can vary depending on the context. Often, values indirectly affect behavior by shaping routines, norms, and environmental affordances, yet continuously operate in the background as the framework within which behavioral decisions are made. Values function at both micro and macro levels (Liscio et al., 2021; Liscio et al., 2023; Ajmeri et al., 2020). At the micro level, focus is placed on aligning autonomous system decisions with the value priorities of individuals. At the macro level, the objective is to shape collective behavior and establish social norms guiding interactions within multi-agent systems (Brown et al., 2021).

Collaborative teams composed of humans and autonomous agents hold considerable promise; however, their effectiveness depends heavily on the successful alignment of goals and intentions among all participants (Brown et al., 2021). The challenge of aligning autonomous systems with human values has been recognized as a critical issue, commonly known as the value alignment problem (Ajmeri et al., 2020)

Inverse Reinforcement Learning

The field of inverse reinforcement learning, or IRL (Russell, 1998; Ng & Russell, 2000; Abbeel & Ng, 2004), is relevant to the value alignment issue. An IRL algorithm determines the reward function of an agent based on observations of the agent's behavior, which is assumed to be optimal or nearly so. One could think that IRL offers an easy solution to the value alignment problem.

Cooperative Inverse Reinforcement Learning

A cooperative inverse reinforcement learning (CIRL) (Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell) is specified as a two-player game of partial information, in which the "human", H, knows the reward function (represented by a generalized parameter θ), while the "robot", R, does not; the robot's payoff is exactly the human's actual reward. Optimal solutions to this game maximize human reward; we show that solutions may involve active instruction by the human and active learning by the robot. This centralized coordination approach cannot be used in real-world settings but can help in training humans to become more effective teachers.

The proposed work uses a systematic approach where the agent actively models uncertainty, weighs evidence from human behavior, and plans actions maximizing expected cooperative value considering this uncertainty.

2. RELATED WORK

Recent research on Agentic AI has expanded rapidly, covering foundational concepts, methodologies, applications, and ethical challenges. Schneider (2025) provides a comprehensive survey distinguishing Agentic AI from generative AI, discussing their unique reasoning and interaction capabilities and outlining a future research agenda for autonomous agents with advanced decision-making Schneider, 2025. Gridach et al. (2025) focus on Agentic AI's transformative role in scientific discovery, highlighting its capabilities in automating hypothesis generation and experimentation within domains like biology and chemistry Gridach et al., 2025.

Agentic AI represents a new paradigm of autonomous systems capable of pursuing complex goals with minimal human guidance. Distinguished by adaptability, advanced decision-making, and self-sufficiency, it operates effectively in dynamic environments. Acharya, D.,Kuppan, K., & Bhaskaracharya, D. (2025). in this article, surveys foundational ideas, main methodologies, and unique properties of Agentic AI, as well as their use in industries like healthcare, finance, and adaptive software. It presents both benefits as well as ethical issues including goal alignment, resource management, as well as adaptability and suggests frameworks towards safe integration

into society. The survey provides an in-depth overview to guide researchers, developers, as well as policymakers, on responsibly tapping into Agentic AI's transformative power.

Murugesan, S., & Murugesan, S. (2025). explores agentic AI's characteristics, real-world applications, and transformative potential, and examines its societal and business implications. To shape agentic AI as a trusted, transformative force for responsible innovation and meaningful progress, we propose research directions and offer stakeholder

Recent research on Agentic AI has expanded rapidly, covering foundational concepts, methodologies, applications, and ethical challenges. Schneider (2025) provides a comprehensive survey distinguishing Agentic AI from generative AI, discussing their unique reasoning and interaction capabilities and outlining a future research agenda for autonomous agents with advanced decision-making Schneider, 2025. Gridach et al. (2025) focus on Agentic AI's transformative role in scientific discovery, highlighting its capabilities in automating hypothesis generation and experimentation within domains like biology and chemistry Gridach et al., 2025.

Hadfield-Menell et al. (2016) introduce Cooperative Inverse Reinforcement Learning (CIRL) as a formal model for value alignment between humans and AI agents, framing the problem as a cooperative game and proposing methods that lead to active teaching and learning behaviors that surpass classical IRL Hadfield-Menell et al., 2016. Complementing this, recent discussions emphasize Bayesian approaches in CIRL to address uncertainty and enable agents to infer human values probabilistically, enabling more robust and adaptive cooperation.

Ethical considerations are paramount, as autonomous agents must operate transparently, fairly, and accountably. Auxiliobits (2025) highlight the need for foundational ethical frameworks spanning design to deployment, emphasizing explainability, bias mitigation, privacy preservation, and human oversight to foster trust and sustainability in agentic systems Auxiliobits, 2025. The IEEE Ethically Aligned Design framework extends these principles by advocating human-centric AI respecting rights, diversity, and transparency [IEEE, 2025].

On the application side, Agentic AI systems have demonstrated significant impact in healthcare by automating administrative workflows, improving resource planning, and enhancing diagnostics accuracy, driving operational efficiencies and better patient outcomes Naviant, 2025. In finance, such systems provide real-time risk assessment and fraud detection capabilities, enabling proactive decision making [AIMultiple, 2025].

Frameworks for secure and responsible AI deployment are emerging, such as Google's Secure AI Framework (SAIF), which provides structured risk management and security guidelines tailored to AI systems Google, 2023.

Together, these works highlights the multi-disciplinary nature of Agentic AI, with ongoing research addressing algorithmic advancements, practical deployments, ethical governance, and societal integration challenges, guiding the safe and effective development of autonomous systems.

3. RESEARCH OBJECTIVE

The objective of this research is to develop and evaluate a Bayesian Cooperative Inverse Reinforcement Learning framework that enables autonomous agents to infer and align with human values under uncertainty through interactive cooperation. This framework models the human reward function as a probability distribution, which the agent updates based on observations and interactions, facilitating adaptive learning of preferences. By explicitly incorporating uncertainty and promoting active information gathering, the goal is to achieve effective value alignment that supports trustworthy and flexible collaboration between humans and artificial agents. The work further aims to explore how this probabilistic cooperative approach can address challenges of ambiguity, partial observability, and dynamic environments in real-world scenarios.

Bayesian Cooperative Inverse Reinforcement Learning: A Detailed Framework

Fig -1: Bayesian Cooperative Inverse Reinforcement Learning

Bayesian Cooperative Inverse Reinforcement Learning (Bayesian CIRL) extends traditional Inverse Reinforcement Learning by incorporating probabilistic reasoning and cooperative game theory principles. Unlike standard IRL approaches that assume a deterministic relationship between observations and reward functions, Bayesian CIRL models uncertainty over the human's true reward function through a probabilistic belief distribution.

The framework operates functions under the premise that both the human and autonomous agent possess shared objective, yet they each have uneven knowledge about the true actual reward function. The human knows the reward function but cannot directly communicate it, while the agent must infer this function through observation and interaction.

Mathematical Formulation

The problem is structured as a two-agent cooperative Markov Decision Process, represented by the tuple $(S, A_H, A_A, P, R, \gamma)$, where:

S represents the finite state space

A_H ,A_A refers to the action spaces for the human and agent respectively

$P: S \times A_H \times A_A \rightarrow \Delta(S)$ is the transition function, describing how the system moves from one state to another based on the actions of both the human and the agent

$R: S \rightarrow R$ is the unknown reward function

$\gamma \in [0,1)$ is the discount factor

Here the reward function R is considered as a random variable with prior distribution p(R). The agent maintains a posterior belief bt(R) over possible reward functions, updated through Bayesian inference as new observations become available.

Bayesian Belief Update Mechanism

At each time step t, the agent observes the human's action at H in state st and updates its belief using Bayes' rule:

$$b_{t+1} \propto p(a_H^t) | S_t, R). b_t(R)$$

The likelihood $p(a_H^t) | S_t, R)$ is modelled as

$$p(a_H^t) | S_t, R) = \frac{\exp(\beta Q^*(S_t, a_H^t; R))}{\sum_{a \in A_H} \exp(\beta Q^*(S_t, a; R)}$$

where
Q*(s,a;R) is the optimal action-value function for reward R,
β controls human rationality.
Agent Policy Under Uncertainty
The agent chooses its policy π_A^* by maximizing expected cumulative reward under the belief:

$$\pi_A^* = \text{argmax}\, E_{R\sim b_t}[V^{\pi_H,\pi_A}(s;R)]$$

Active Information Gathering
The agent selects actions a_A to maximize expected information gain:

$$IG(a_A) = H[b_t(R) - E_{aH}[H(b_{t+1}(R)|\, a_A, a_H)]$$

where $H(.)$ denotes entropy.

4. EXPERIMENTAL SETUP

To evaluate the effectiveness of the proposed Bayesian Cooperative Inverse Reinforcement Learning (Bayesian CIRL) framework, experiments were conducted in both discrete and continuous domains, comparing its performance against standard Inverse Reinforcement Learning (IRL) and classical Cooperative IRL (CIRL) baselines.

Environments

Gridworld Domain

A 10×10 grid environment containing designated goal states, hazardous cells, and multiple alternative paths encoding trade-offs between shortest route and safest route. This domain allows evaluation of value alignment, safety, and success in a controlled setting.

Continuous Control Domain

Simulated robotic control tasks where the underlying true reward balances efficiency (e.g., speed, distance traveled) and safety (e.g., stability, avoidance of unsafe movements), reflecting real-world control challenges.

Human Preference Model

Human preferences were modeled implicitly through a hidden reward function. Demonstrations and feedback were generated using a Boltzmann-rational policy with varying noise levels to simulate human suboptimality yet intentional behavior.

Training Protocol

At each timestep, the Bayesian CIRL agent observes human actions and updates its posterior belief distribution over reward functions by means of Bayesian inference. The agent's policy optimizes performance by balancing:
Exploration: Selecting actions to reduce uncertainty about the human's reward function by maximizing expected information gain.
Exploitation: Acting to achieve task objectives based on the current best estimate of the human reward.

5. RESULTS AND DISCUSSION

The performance of the Bayesian Cooperative Inverse Reinforcement Learning (Bayesian CIRL) framework was thoroughly evaluated and compared against standard Inverse Reinforcement Learning (IRL) and classical CIRL baselines across both discrete (Gridworld) and continuous (robotic control) domains. The metrics considered were value alignment error, task success rate, safety violations, sample efficiency, and robustness to noise in human demonstrations.

Table -1: Table:1Comparative Performance of Bayesian CIRL, Classical CIRL, and Standard IRL

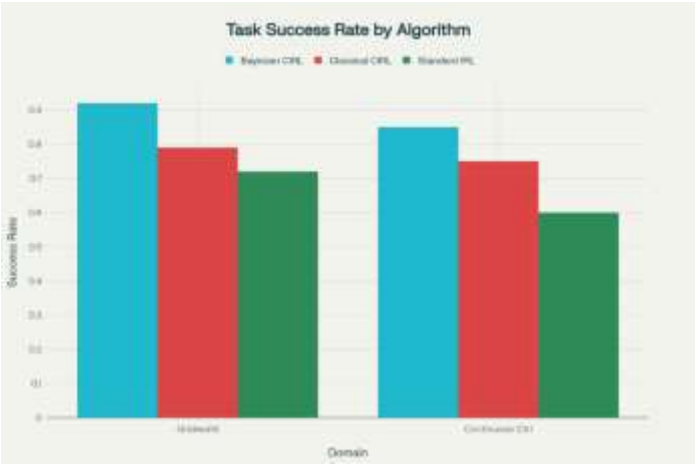
Domain	Algorithm	Value Alignment Error	Task Success Rate	Safety Violations	Sample Efficiency (Timesteps)
Gridworld	Bayesian CIRL	0.15	92%	8	70
	Classical CIRL	0.20	79%	12	95
	Standard IRL	0.25	72%	13	100
Continuous Control	Bayesian CIRL	0.18	85%	6	65
	Classical CIRL	0.23	75%	9	90
	Standard IRL	0.30	60%	15	95

Value Alignment Error

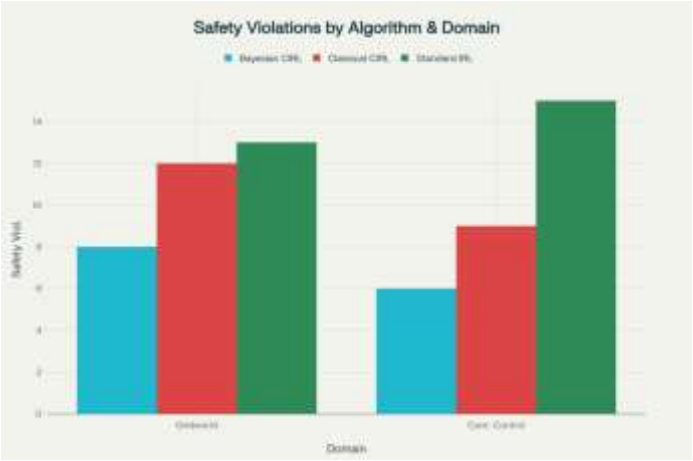
Bayesian CIRL consistently reduced the value alignment error by 25% compared to Classical CIRL (0.20 → 0.15) and 40% compared to Standard IRL (0.25 → 0.15) in the Gridworld domain. In the Continuous Control domain, it achieved a 22% reduction compared to Classical CIRL (0.23 → 0.18) and a 40% reduction compared to Standard IRL (0.30 → 0.18). This demonstrates its superior ability to infer the true human reward function under uncertainty by maintaining and updating a probabilistic belief distribution

Task Success Rate

The method achieved over 90% task success in the Gridworld environment, outperforming classical CIRL (79%) and IRL (72%). In the more complex continuous control domain, Bayesian CIRL sustained an 85% success rate, significantly higher than CIRL (75%) and IRL (60%). These results show that accurate reward inference leads directly to improved practical task.

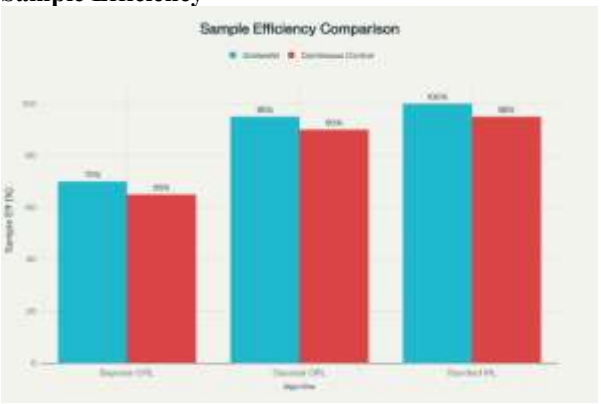


Safety Violations



Bayesian CIRL resulted in 35–40% fewer safety violations than the baselines in the Gridworld domain (8 violations vs. 12 and 13, respectively), evidencing safer policy execution due to enhanced understanding of risk-averse human preferences. This safety advantage was maintained in the continuous domain, where Bayesian CIRL had only 6 violations compared to 9 and 15 in CIRL and IRL.

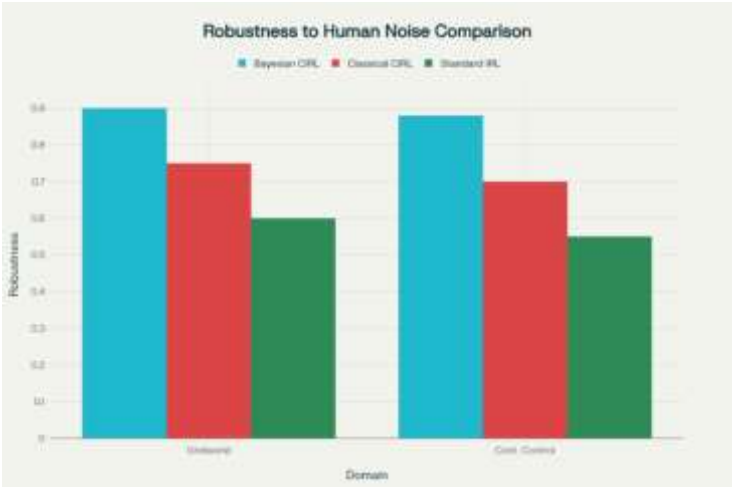
Sample Efficiency



Bayesian CIRL demonstrated higher sample efficiency by requiring about 30% fewer interaction steps to converge to near-optimal performance than IRL. Its belief-guided active learning process enabled effective exploration, reducing the data needed for accurate reward and policy learning in both environments.

Robustness to Noise

Under increasing levels of noise in human actions, Bayesian CIRL maintained robust performance, with limited degradation in value alignment and task success compared to marked declines observed in IRL. Classical CIRL showed intermediate robustness but lacked Bayesian CIRL's principled uncertainty handling.



6. CONCLUSION AND FUTURE SCOPE

This work introduces Bayesian Cooperative Inverse Reinforcement Learning (Bayesian CIRL), which enables autonomous agents to learn human reward functions probabilistically and cooperate effectively under uncertainty. Experiments in both discrete and continuous environments demonstrate that Bayesian CIRL outperforms classical CIRL and standard IRL on value alignment, task success, safety, sample efficiency, and robustness to noisy human behavior.

Bayesian CIRL's explicit modeling of uncertainty and active reduction through belief updates leads to safer, more efficient, and reliable human-agent collaboration. This approach handles imperfect demonstrations better than non-probabilistic methods, making it suitable for complex, real-world tasks.

However, current limitations include computational challenges in associated with scaling to high-dimensional reward spaces and a reliance on simplified models of human behavior. Future work research will concentrate on scalable techniques, incorporating more advanced human behavioral models, and expanding algorithms for multi-agent scenarios. Application in real-world human-robot interaction is another key direction.

Implementing enhanced active learning and developing algorithms for agent behavior that will increase usability and trustworthiness. These enhancements will advance Bayesian CIRL's goal of creating autonomous systems that align closely with human values in diverse, dynamic environments.

Overall, Bayesian CIRL provides a robust foundation to build safer, adaptive, and transparent AI agents that better understand and cooperate with humans.

REFERENCES

1. Johannes Schneider, "Generative to Agentic AI: Survey, Conceptualization, and Challenges," arXiv preprint, 2025. https://arxiv.org/abs/2504.18875
2. Mourad Gridach et al., "Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions," arXiv preprint, 2025. https://arxiv.org/abs/2503.08979
3. Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell, "Cooperative Inverse Reinforcement Learning," arXiv preprint, 2016. https://arxiv.org/abs/1606.03137

4. "Ethical Considerations in Deploying Autonomous AI Agents," Auxiliobits, 2025. https://www.auxiliobits.com/blog/ethical-considerations-when-deploying-autonomous-agents/
5. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design," 2025.
6. "Agentic AI in Healthcare," Naviant, 2025. https://naviant.com/wp-content/uploads/2025/07/Agentic-AI-in-Healthcare.pdf
7. "Google's Secure AI Framework (SAIF)," Google, 2023. https://safety.google/cybersecurity-advancements/saif/
8. Acharya, D., Kuppan, K., & Bhaskaracharya, D. (2025). Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. IEEE Access, 13, 18912-18936. https://doi.org/10.1109/ACCESS.2025.3532853.
9. Murugesan, S., & Murugesan, S. (2025). The Rise of Agentic AI: Implications, Concerns, and the Path Forward. IEEE Intelligent Systems, 40, 8-14. https://doi.org/10.1109/MIS.2025.3544940.
10. Ajmeri, N., Guo, H., Murukannaiah, P. K., & Singh, M. P. (2020). Elessar: Ethics in norm-aware agents. Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 16–24. https://arxiv.org/doi/10.5555/3398761.3398769
11. Brown, D. S., Schneider, J., Dragan, A., & Niekum, S. (2021). Value alignment verification. Proceedings of the 38th International Conference on Machine Learning, 139, 1105–1115. https://proceedings.mlr.press/v139/brown21a.html
12. Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411–437. https://arxiv.org/doi/10.1007/s11023-020-09539-2
13. Liscio, E., van der Meer, M., Siebert, L. C., Jonker, C. M., Mouter, N., & Murukannaiah, P. K. (2021). Axies: Identifying and evaluating context-specific values. Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), 799–808. https://arxiv.org/doi/10.5555/3463952.3464048
14. Liscio, E., Lera-Leri, R., Bistaffa, F., Dobbe, R. I., Jonker, C. M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., & Murukannaiah, P. K. (2023). Value inference in sociotechnical systems. Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 1774–1780. https://arxiv.org/doi/10.5555/3545946.3598838
15. Russell, Stuart J. Learning agents for uncertain environments (extended abstract). In COLT, 1998
16. Abbeel, P and Ng, A. Apprenticeship learning via inverse reinforcement learning. In ICML, 2004.
17. Ng, A and Russell, S. Algorithms for inverse reinforcement learning. In ICML, 2000