Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Artificial Intelligence Value Alignment via Inverse Reinforcement Learning

João L. Duim[1] Diego P. P. Mesquita[2]
EMAp/FGV, Rio de Janeiro, RJ

Value alignment, one of the Artificial Intelligence (AI) Alignment problems, pertains to ensuring that AI systems adhere to human values. These intricate problems lack definitive solutions, and significant research has been conducted to address it [3]. Nevertheless, substantial progress is still required to effectively tackle the AI Alignment problems. Current trajectories in AI development, particularly in the realm of Deep Learning and RLHF [4], pose significant existential risks due to potential misalignments in AI objectives and human values.

In the present study, our objective is to address the AI value alignment problem through the utilization of Inverse Reinforcement Learning (IRL) [1]. The central idea revolves around employing the IRL framework to acquire a reward function from an expert who exhibits behaviour consistent with human values. Subsequently, the AI system will mimic the expert's actions, thereby aligning its behaviour with human values in a verifiable way [2].

Definition 1: [MDP] Let S be a finite set of states, S_0 be a distribution over initial states, A a set of actions, $T : S \times A \to \mathcal{P}(S)$ a transition probability distribution, $\gamma \in [0, 1]$ a discount factor and $R : S \times A \to \mathbb{R}$ a reward function. $M := (S, A, T, S_0, \gamma, R)$ is a Markov Decision Process and $M \backslash R = (S, A, T, S_0, \gamma)$ is said to be the environment.

Definition 2: [Policy] Given an MDP M, a (stochastic) policy $\pi : S \to \mathcal{P}(A)$ is a probability distribution over the next action choice given the current state.

Definition 3: [State-action value] Given an MDP M and a policy π, the state-action value function $Q^\pi_R : S \times A \to \mathbb{R}$ is given by

$$Q^\pi_R(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t)) \mid s_0 = s, a_0 = a \right]. \tag{1}$$

Definition 4: [Optimality] The optimal policy π^* satisfies $Q^{\pi^*}_R(s, a) = \sup_\pi Q^\pi_R(s, a)$, $\forall (s, a) \in S \times A$. Solving an MDP consists of learning an optimal policy, and a way to do that online is Reinforcement Learning.

Definition 5: [IRL] Assume that there is an agent called *learner* and another agent called *expert* behaving according to an underlying policy π_E, which may not be known. Let an MDP $M \backslash R_E$ model the expert behaviour. Let $D = \{\langle (s_0, a_0), (s_1, a_1), ..., (s_j, a_j) \rangle_1, \langle (s_0, a_0), (s_1, a_1), ..., (s_j, a_j) \rangle_2, ..., \langle (s_0, a_0), (s_1, a_1), ..., (s_j, a_j) \rangle_N\}$ be the set of demonstrated trajectories, all of them perfectly observed. Then, determine $\hat{R}_E$ that best explains either policy π_E if given or the observed behaviour in the form of demonstrated trajectories.

[1] jlduim@gmail.com
[2] diego.mesquita@fgv.br

Definition 6: [ϵ-value alignment] Let $M := (S, A, T, S_0, \gamma, R)$ be the MDP that models the sequential decision-making of an agent. A policy π' is said to be ϵ-value aligned in environment $M \backslash R$ if and only if

$$Q^*_R(s,a) - Q^{\pi'}_R(s,a) \leq \epsilon, \forall (s,a) \in S \times A \tag{2}$$

Exact value alignment is achieved when $\epsilon = 0$.

Definition 7: [OPT] Given an agent behaving as MDP M, $OPT(R) = \{\pi \mid \pi(a \mid s) > 0 \Longrightarrow a \in \arg\max_a Q^*_R(s,a)\}$ is defined as the set of all optimal policies.

Corollary 1: Exact value alignment between expert and learner in environment E is achieved if $OPT(R') \subseteq OPT(R)$.

Definition 8: [CRS] In environment E, the consistent reward set of a policy π is defined as $CRS(\pi) = \{R \mid \pi \in OPT(R)\}$, which is the set of reward functions under which π is optimal.

Theorem 1: Efficient exact value alignment verification is possible in the following query settings: 1) Query access to reward function weights $\boldsymbol{w}'$; 2) Query access to samples of the reward function $R'(s)$; 3) Query access to $V^*_{R'}(s)$ and $Q^*_{R'}(s,a)$; 4) Query access to preference over trajectories.

Definition 9: [ARS] The aligned reward set is defined as $ARS(R) = \{R' \mid OPT(R') \subseteq OPT(R)\}$.

Theorem 2: Assuming $R(s) = \boldsymbol{w}^T\phi(s)$ and $R'(s) = \boldsymbol{w'}^T\phi(s)$, where $\phi(s) \in \mathbb{R}^k$, and given an optimal policy π^*_R under R then

$$\boldsymbol{w}' \in \bigcap_{(s,a,b)\in O} \mathcal{H}^R_{s,a,b} \Longrightarrow R' \in ARS(R) \tag{3}$$

where $\mathcal{H}^R_{s,a,b} = \{\boldsymbol{w} \mid \boldsymbol{w}^T(\Phi^{(s,a)}_\pi - \Phi^{(s,b)}_\pi) > 0\}$ and $O = \{(s,a,b) \mid s \in S, a \in A(s), b \notin A(s)\}$.

Theorem 2 provides sufficient condition for verifying exact value alignment.

Acknowledgements

Thankful to my advisor, FGV CDMC, FGV EMAp, Condor Initiative and Effective Thesis — for their collective role in shaping my academic journey.

References

[1] S. Arora and P. Doshi. "A survey of inverse reinforcement learning: Challenges, methods and progress". In: **Artificial Intelligence** 297 (2021), p. 103500.

[2] D. S. Brown, J. Schneider, A. Dragan, and S. Niekum. "Value alignment verification". In: **International Conference on Machine Learning**. PMLR. 2021, pp. 1105–1115.

[3] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, et al. "AI Alignment: A Comprehensive Survey". In: **arXiv preprint arXiv:2310.19852** (2023).

[4] R. Ngo, L. Chan, and S. Mindermann. "The alignment problem from a deep learning perspective". In: **arXiv preprint arXiv:2209.00626** (2022).