Social Impact Regularization in IQ-Learn

Steering Social Intent in Heterogeneous Driving Demonstrations

by

Petar Koev

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday, May 28, 2026 at 3:00 PM.

Student number:	5466539	
Project duration:	October 1, 2025 – May 28, 2026	
Thesis committee:	Dr. L. Cavalcante Siebert,	TU Delft, Thesis Advisor
	Antonio Mone,	TU Delft, Daily Co-Supervisor
	Dr. Chirag Raman,	TU Delft, Thesis Committee Member

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Summary

Autonomous driving relies heavily on Reinforcement Learning (RL) to train agents in sequential decision-making settings. However, RL's success is deeply bottlenecked by the need to manually specify a reward function, a notoriously difficult task when attempting to balance safety, efficiency, and nuanced social etiquette in highly interactive domains. Inverse Reinforcement Learning (IRL) circumvents this challenge by extracting latent objectives directly from expert data. Yet, standard IRL operates under a critical assumption: that all demonstrations stem from a single, homogeneous behavioural profile. In reality, traffic is fundamentally heterogeneous, composed of a mixture of distinct driving styles ranging from calm and cooperative to aggressive and assertive. When standard IRL is applied to such mixed datasets, it inherently struggles to fit a single reward function to the conflicting behaviours. Consequently, the recovered reward typically collapses into an arbitrary average, completely misrepresenting varied driving profiles and failing to account for the essential social context of driving. To resolve this ambiguity, this thesis introduces the Social Impact Regularized IQ-Learn framework. This approach decomposes the driving reward into two distinct components: an individual reward capturing the ego vehicle's own progress, and an ego-centric social impact signal measuring how the vehicle's actions directly affect its neighbours. By combining these into a social scoring function, the framework integrates a normative prior as an additive regularizer within the IQ-Learn objective. This formulation exploits a vital separation: the core IQ-Learn objective absorbs universal physical driving dynamics from the entire mixed dataset, while the regularizer selectively steers the social interpretation of those dynamics towards a specific, designer chosen behavioural target. Evaluations spanning a tabular gridworld proof of concept, a multi-agent stochastic environment, and a continuous observation intersection simulator confirm that the regularizer effectively resolves behavioural ambiguity. The framework can successfully steer the recovered policy towards a targeted social alignment. Ultimately, by making the social orientation of the learned policy an explicit and inspectable parameter, this methodology provides a concrete, auditable mechanism for designers and regulators to verify that an autonomous vehicle's social behaviour actively matches its intended design.

Contents

1

Introduction

Autonomous driving is not purely a control task, it is fundamentally interactive [1, 2]. Driving scenarios involve constant negotiations, from merging into tight gaps to navigating ambiguous intersections and resolving right-of-way disputes. These interactions are rarely communicated through explicit signaling. Instead, they depend on drivers inferring each other's intent through implicit communication and adjusting accordingly to ensure safety [3]. Prior work has framed such situations as social dilemmas, where individual goals and collective cooperation continuously clash [4]. Capturing this within a decision-making framework is a central challenge for autonomous driving.

Reinforcement Learning (RL) offers a powerful tool for training autonomous agents in such sequential decision making settings. Its success, however, relies entirely on the manual specification of a reward function [5, 6]. In highly interactive domains, hand crafting a reward that balances safety, efficiency and nuanced social etiquette is notoriously difficult [7]. System designers often struggle with incomplete specifications, resulting in agents that fail to capture the subtle notion of human driving or exploit mathematical loopholes to find unintended solutions that completely miss the core objective of the driving task. To bypass this challenge, researchers have turned to extracting these objectives directly from data.

Inverse Reinforcement Learning (IRL) attempts to recover latent reward functions from demonstrations [8, 9], aiming to model implicit preferences such as safety, comfort and efficiency [10, 11]. When applied to driving data, however, standard IRL typically operates under a critical assumption: that all demonstrations come from a single homogeneous reward function [12, 13, 14]. In practice traffic consists of a mixture of driving behaviours, such as calm, aggressive, cooperative or assertive styles. When traditional IRL is applied to such heterogeneous data, it struggles by attempting to fit a single reward function to inherently conflicting behaviours. Rather than capturing the distinct underlying intents, the recovered reward typically collapses into an arbitrary average or disproportionately favours a single behavioural mode present in the dataset [8, 12, 13]. This single reward approach not only misrepresents the varied driving profiles but entirely fails to account for the social context of driving. All of the interactive, cooperative or even competitive dynamics that govern real world driving are lost.

While Multi-Intent IRL (MI-IRL) provides a potential mechanism to address this heterogeneity by explicitly learning multiple reward functions, it introduces significant computational complexity [13, 15]. A more practical approach is to explicitly encode sociality into a unified framework. Social Value Orientation (SVO), originating in social psychology [16], provides a geometric parametrization of an agent's trade-off between individual and collective outcomes via a single angle α. In autonomous driving, SVO has been applied to game-theoretic forward planners parametrized by the estimated social angle of surrounding drivers [17, 18] and in its classical formulation to inverse reinforcement learning for inferring the latent preferences of interacting drivers [19]. In these settings, the social angle is either an

input to the planner or a latent parameter estimated per driver alongside the reward. Because standard frameworks rarely formalize this as a prior to regularize reward functions across heterogeneous datasets, policies learned from mixed driving data often learn what a car can do, rather than what a car should do. Such frameworks lack principled ways to distinguish socially meaningful choices, like waiting for a pedestrian, from any other kinematically feasible alternatives such as trajectories that successfully avoid crashes but completely ignore human driving etiquette.

The challenge, therefore, is not just to model the full spectrum of driving behaviour, but to selectively filter for socially desirable modes when the dataset contains heterogeneous intentions. Modern driving datasets aggregate trajectories from drivers with fundamentally different social preferences. Some prioritize their own progress, others yield to surrounding traffic, and the proportions are unknown [14, 20, 21]. Since IRL yields a degenerate set of valid reward functions from mixed data, the goal is no longer to hand craft the complete objective. Instead, it is to leverage a structural prior that enforces a desired social orientation. This allows the system designer to select a target behavioural alignment, ensuring that the resulting policy follows specific norms without compromising the universal driving capabilities learned from the expert data.

Recent work in inverse reinforcement learning, namely IQ-Learn [22], offers a robust and scalable foundation for this objective. It recovers an implicit reward function directly from a single soft Q-function, entirely avoiding the brittle adversarial inner loop of classical maximum entropy IRL. Building upon this stable architecture to resolve the ambiguity of mixed demonstration data, we propose the Social Impact Regularized IQ-Learn framework. First, we decompose the driving reward into two hand-specified components: an individual reward capturing the ego's own progress and an ego-centric social impact signal that measures how the ego vehicle's actions affect its neighbours. This decomposition yields distinct behavioural modes that correspond to distinguishable points in the reward space. Second, we incorporate this social signal as a regularizer on the reward function recovered by IQ-Learn [22], steering it toward the target social behaviour specified by the designer. This is motivated by two concrete failure modes that arise when IQ-Learn is trained on heterogeneous data. In ambiguous datasets, where the data contains equal proportions of conflicting behavioural modes, the recovered reward can arbitrarily collapse onto one mode or oscillate between them. In contaminated datasets, where a majority of safe demonstrations is mixed with a minority of adversarial trajectories, the recovered reward partially absorbs the adversarial intent, increasing collision rates above what an individual safe expert would produce. The regularizer exploits a vital separation to address both failure modes: the IQ-Learn objective absorbs universal physical driving dynamics from the entire dataset, while the regularizer selectively steers the social interpretation of those dynamics toward a specific, designer-chosen behavioural target.

1.1. Research Questions and Contributions

The proposed Social Impact Regularized IQ-Learn framework serves as the methodological foundation of this work, developed to address the following research questions:

- How does social impact regularization influence the resolution of behavioural ambiguity in multimodal demonstration datasets, and to what extent does it enable the selection of a specific behavioural mode?
- To what extent can the regularizer steer the recovered reward function toward a specified social profile, and what are the trade-offs between social alignment and driving competence?

To answer these questions, this thesis makes the following contributions to the field of socially aware autonomous driving:

1. We propose a method to integrate an ego-centric social prior as an additive regularizer into the IQ-Learn objective.
2. We define a social signal as a measure of how the ego vehicle's actions affect its neighbours, rather than as the aggregate welfare of other agents.
3. We validate the framework on two tabular gridworlds where optimal policies are known exactly and on a continuous observation intersection simulator with reactive traffic.

The implications of this problem extend beyond the technical. Autonomous vehicles are no longer a research prototype, as commercial deployments are active, and regulatory frameworks governing AI in safety critical domains are rapidly being formalized. A system whose social alignment is an unintended artifact of its training data, rather than a deliberate design choice, is difficult to certify and even harder to hold accountable when its behaviour causes harm. By making the social orientation of the learned policy an explicit, inspectable parameter, the proposed framework provides a concrete mechanism for this accountability, enabling designers and regulators to verify that a vehicle's social behaviour matches its stated intent.

The remainder of this thesis is organized as follows. Section 2 introduces the necessary background knowledge needed, covering the Markov Decision Process formulation, inverse reinforcement learning, and the IQ-Learn framework. Section 3 reviews related work on inverse reinforcement learning in autonomous driving, social metrics for behavioural modelling and use of regularization in inverse reinforcement learning methods. Section 4 presents the proposed methodology, including the social reward decomposition and the mathematical analysis of the integration strategy. Section 5 describes the experimental setup. Section 6 presents the results. Section 7 interprets the findings and discusses their broader implications. Section 8 discusses limitations and directions for future work. Section 9 reflects on the social implications of the proposed framework. Finally, Section 10 concludes the thesis.

2

Background

This chapter establishes the theoretical foundations upon which the proposed framework is built. We begin by formally defining the autonomous driving problem as a Markov Decision Process, providing the standard mathematical structure for sequential decision making. Following this, we introduce Inverse Reinforcement Learning and the specific IQ-Learn architecture used to recover latent objectives from expert demonstrations. Finally, we outline Tikhonov regularization, detailing the mathematical mechanism that allows explicit priors to be incorporated into otherwise ill-posed inverse problems.

2.1. Markov Decision Process

The autonomous driving problem is defined as a continuous Markov Decision Process (MDP). The MDP is defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, \gamma \rangle$. The state space $\mathcal{S} \in \mathbb{R}^n$ captures the ego vehicle's kinematic state together with the relevant features of surrounding traffic participants. The action space $\mathcal{A} \in \mathbb{R}^m$ consists of continuous control inputs, specifically longitudinal acceleration and steering commands. The transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ models the deterministic dynamics of the vehicle and environment. The discount factor $\gamma \in [0, 1)$ governs the agent's effective planning horizon by weighting immediate versus future rewards.

The agent aims to learn a policy $\pi(a \mid s)$ that maximizes the expected discounted return:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \, R(s_t, a_t) \right]. \tag{2.1}$$

In standard driving formulations, the reward function $R(s, a)$ is typically treated as a fixed specification, encoding objectives such as safety, efficiency, and comfort. In our setup, we adopt this conventional formulation for the MDP definition. The departure from a fixed reward and the incorporation of social intent is introduced later, through the IRL framework.

2.2. Inverse Reinforcement Learning

In many settings, and autonomous driving in particular, the reward function $R(s, a)$ is not available as a fixed specification. Instead, the agent has access to a dataset of expert demonstrations $\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$, where each trajectory $\tau = (s_0, a_0, s_1, a_1, \ldots)$ is assumed to have been generated by an expert policy π_E that is (near-)optimal under some unknown reward $R^\star$. Inverse Reinforcement Learning (IRL) addresses the problem of recovering $R^\star$, or a behaviourally equivalent reward, from $\mathcal{D}$ [8].

Formally, IRL seeks a reward R such that the expert policy π_E is optimal with respect to it:

$$\mathbb{E}_{\pi_E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right] \;\geq\; \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right] \qquad \forall \, \pi. \tag{2.2}$$

Because Equation (2.2) alone admits infinitely many solutions (including the trivial $R \equiv 0$), modern formulations replace the hard optimality constraint with a probabilistic model of the expert. The Maximum Entropy IRL formulation [23] assumes the expert is Boltzmann-rational, so trajectories are observed with probability

$$p(\tau \mid R) \;\propto\; \exp\!\left(\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right), \tag{2.3}$$

and the reward is recovered by maximising the likelihood of the demonstration set $\mathcal{D}$ under this model. Equation (2.3) also serves as the conceptual foundation for the soft-Q formulation of IQ-Learn introduced in Section 2.2.2.

A core challenge of IRL is that the problem is fundamentally ill-posed: many reward functions can explain the same set of demonstrations [8, 15]. While this ambiguity means the recovered reward is always underdetermined, additional constraints provide a principled mechanism to select a specific, functional reward from the set of valid explanations. To resolve this ambiguity, the literature offers two primary mechanisms: modelling the demonstrator population as a mixture of distinct intentions (Section 2.2.1), and reformulating the reward recovery process entirely through an implicit value function (Section 2.2.2).

2.2.1. Multi-Intent Inverse Reinforcement Learning

Standard IRL assumes that all trajectories in $\mathcal{D}$ are generated by a single expert optimising a single reward $R^\star$. In practice, naturalistic driving datasets aggregate trajectories from drivers with systematically different preferences, ranging from aggressive to yielding and cooperative profiles, which inherently violates the single-reward assumption. Multi-Intent IRL (MI-IRL) relaxes this assumption by modelling the dataset as a mixture of K latent expert types, each with its own reward [12, 13]. Each trajectory $\tau_i \in \mathcal{D}$ is associated with a latent assignment $z_i \in \{1, \dots, K\}$, and the joint objective recovers a set of rewards $\{R_k\}_{k=1}^K$ together with a distribution over the assignments:

$$\{R_k^\star\}_{k=1}^K,\ \{z_i^\star\}_{i=1}^N \;=\; \arg\max_{\{R_k\},\,\{z_i\}} \prod_{i=1}^{N} p\big(\tau_i \mid R_{z_i}\big)\; p(z_i), \tag{2.4}$$

where $p(\tau \mid R)$ is the likelihood induced by the chosen IRL model (e.g., Equation (2.3)) and $p(z_i)$ is a prior over the latent mode assignment. Inference is typically performed by expectation-maximization or, in the Bayesian variant, by Markov chain Monte Carlo sampling over rewards and assignments [13].

2.2.2. IQ-Learn

IQ-Learn [22] reformulates inverse reinforcement learning as a Q-learning problem, avoiding the need for an explicit discriminator. The key insight is that the reward can be recovered implicitly from the Q-function. Given a soft Q-function $Q(s, a)$ and the corresponding soft value function

$$V(s) \;=\; \tau \cdot \mathrm{logsumexp}(Q(s, \cdot)/\tau)\,, \tag{2.5}$$

the implicit reward for a transition (s, a, s') is defined as

$$\hat{r}(s, a) = Q(s, a) - \gamma(1 - d)V(s'), \tag{2.6}$$

where γ is the discount factor, d is the terminal flag, and τ is a temperature parameter. Rather than learning a separate reward network, IQ-Learn trains Q directly by minimizing a divergence between the occupancy measures of the expert and the learner. The resulting loss consists of three terms:

$$\mathcal{L}_{\mathrm{IQ}} = \underbrace{-\mathbb{E}_{\rho_E}\left[\varphi'(\hat{r}) \cdot \hat{r}\right]}_{\text{Term 1}} + \underbrace{\mathbb{E}_{\rho}\left[V(s) - \gamma V(s')\right]}_{\text{Term 2}} + \underbrace{\frac{1}{4\alpha_\chi}\mathbb{E}_{\rho_E}\left[\hat{r}^2\right]}_{\text{Term 3}} \tag{2.7}$$

Term 1 maximizes the implicit reward on expert transitions, weighted by φ', the derivative of the chosen f-divergence generator. Term 2 regularizes the value function, with the expectation taken over expert data, a mixture of expert and learner data, or initial states, depending on the sampling strategy. Term 3, active when using the χ^2 divergence, bounds the magnitude of the learned implicit reward, stabilizing training. This term acts as a zero-centered Tikhonov regularizer on the recovered reward (Section 2.3).

2.3. Tikhonov Regularization

Tikhonov regularization [24] is the standard method for solving ill-posed inverse problems, where the data alone does not uniquely determine a solution. In the context of Inverse Reinforcement Learning, it provides the mathematical mechanism to resolve this inherent ambiguity and break the symmetry of the solution space. Given a data fidelity objective $\mathcal{L}_{data}(\theta)$, the regularized problem adds a quadratic penalty that biases the solution toward a prior estimate θ_{prior}:

$$\mathcal{L}_{\text{regularized}} = \mathcal{L}_{\text{data}}(\theta) + \lambda \|\theta - \theta_{\text{prior}}\|^2. \tag{2.8}$$

While classically formulated for discrete parameter vectors θ, this concept naturally extends to function spaces. In the context of recovering rewards, it can be applied directly to the recovered implicit reward function $\hat{r}(s, a)$, penalizing its divergence from a baseline prior function. When $\theta_{prior} = 0$, this reduces to standard L_2 regularization, which penalizes large parameter values as seen with the χ^2 penalty in IQ-Learn, described in Section 2.2.2. However, when an informed prior is used ($\theta_{prior} \neq 0$) to encode domain knowledge, the regularizer does more than constrain magnitude, it pulls the solution toward a semantically meaningful target. This allows the system designer to deliberately select, from the degenerate set of valid rewards, the specific functional profile that is closest to the prior estimate. The strength parameter λ controls the trade-off between data fidelity and this prior conformity. While other regularization methods are available, including L_1 sparsity penalties and f-divergences, the quadratic Tikhonov form remains uniquely characterized by its Bayesian equivalence to a Gaussian prior over the function space.

3

Related Work

This chapter positions the proposed framework within the broader landscape of autonomous driving and inverse reinforcement learning literature. It reviews existing approaches to highlight the persistent gaps in handling mixed driving datasets and enforcing normative behaviours. Section 3.1 examines the challenges of learning from heterogeneous demonstrations, contrasting standard and multi-intention approaches. Section 3.2 surveys social metrics and behavioural modelling, establishing the foundation for the ego-centric social impact signal used in this work. Finally, Section 3.3 looks at regularization in reward recovery, demonstrating how the proposed approach builds upon modern implicit regularizers by introducing a semantically informed prior.

3.1. Learning Driving Rewards from Heterogeneous Demonstrations

The standard approach to inverse reinforcement learning (IRL) for driving trains a single reward model on combined demonstrations from large naturalistic datasets, such as NGSIM [25], HighD [26], and the Waymo Open Motion Dataset [27]. This approach implicitly treats the pooled trajectories as if they were drawn from a single expert behaviour [10, 11, 15]. However, naturalistic driving is fundamentally heterogeneous, with drivers differing systematically in aggressiveness, yielding behaviour, and overall social preference [14, 20, 21]. A related line of domain-agnostic work attempts to address trajectory variation by relaxing the optimality assumption to handle imperfect demonstrations. While techniques like Confidence-Aware Imitation Learning [28] weigh state-action pairs and T-REX [29] or D-REX [30] rely on ranked trajectories, these methods ultimately remain ill-suited for driving. Realistically, the explicit quality rankings or optimality labels that such methods rely on are unavailable in mixed driving datasets. More importantly, treating behavioural variation as suboptimality is a conceptual mismatch. Diverse driving styles reflect legitimate preference differences rather than errors to be filtered out.

A distinct class of methods, Multi-Intention IRL (MI-IRL), explicitly models demonstration heterogeneity by clustering trajectories into behavioural modes and recovering a separate reward function for each [12, 13]. While these recovered modes remain interpretable when using handcrafted features, this clarity is lost in recent high-dimensional deep learning extensions [31, 32]. When rewards are parameterized by complex neural networks, the resulting modes often reflect statistical artifacts of the clustering procedure rather than genuine, interpretable behavioural profiles. Furthermore, a structural limitation persists across all MI-IRL methods: the algorithm returns a collection of rewards but provides no intrinsic mechanism for the designer to dictate which should govern the deployed policy. Our framework inverts this paradigm. Rather than discovering behavioural modes through unsupervised clustering, we enable the designer to specify a target social profile upfront using a semantically grounded prior. From the set of reward functions consistent with the demonstration data, our IRL objective directly recovers the single reward that best matches this specified target.

3.2. Social Metrics and Behavioural Modelling

Several social metrics have been proposed for autonomous driving, each capturing a different aspect of interaction. Inequity Aversion (IA) frameworks model traffic as non-cooperative games where agents seek Nash Equilibria [2], while causal responsibility metrics such as Feasible Action Space Reduction (FeAR) [33] and Responsibility-Sensitive Safety (RSS) [34] quantify how an agent's actions restrict the options available to others. These methods are effective for planning and safety verification, but they either model strategic influence or define negative constraints on what an agent should not do, rather than capturing positive social intentions.

Social Value Orientation (SVO) provides a formal framework for quantifying social preference on a continuous scale. Originating in social psychology [16], SVO projects an agent's preference between personal and collective reward onto the unit circle via an angle α. [19] integrated SVO into game-theoretic planning for merging scenarios, and subsequent work has utilized SVO in multi-agent reinforcement learning to promote highway coordination [35]. SVO is widely used for forward policy generation, but to the best of our knowledge, utilizing a social signal as an IRL regularizer to untangle mixed demonstrations remains unexplored. After careful inspections, however, we depart from the classical SVO definition of the global reward, which aggregates the independent utilities of other agents, and instead define an ego-centric social impact signal that measures how the ego's actions affect its neighbours. This distinction is critical for achieving meaningful mode separation in the regularization target as shown in Appendix A.

3.3. Regularization in Inverse Reinforcement Learning

The reward ambiguity inherent in IRL [8] has been addressed through various forms of regularization. Maximum Entropy IRL [23] selects the reward under which the expert policy has maximum entropy, imposing a smoothness prior. Bayesian IRL [36] provides a more explicit mechanism by placing a prior distribution over reward functions and computing the posterior given the demonstrations. While principled, its primary limitation is the computational expense of posterior inference, which requires sampling over the space of reward functions and solving the forward planning problem for each sample, making it practically infeasible in continuous domains [13, 15]. Additionally, the priors used in practice are typically uninformative (e.g., zero-centered Gaussians), providing regularization but no semantic guidance.

Beyond smoothness and uninformative Gaussian priors, other approaches enforce sparsity using L_1 regularization, operating under the assumption that experts base their decisions on a small subset of available features. While effective for feature selection and producing minimal, interpretable reward functions, sparsity priors face the same fundamental limitation: they lack semantic direction. They eliminate variables rather than guiding the recovered reward toward a specific behavioural interpretation.

More recently, advancements in offline and soft-Q IRL, most notably the IQ-Learn framework [22], have introduced implicit regularization through f-divergences. While the χ^2 divergence used in IQ-Learn acts as a zero-centered penalty to stabilize neural network training, it remains a purely mathematical stabilizer. It prevents the reward from collapsing or overfitting, but it provides no behavioural guidance to untangle the conflicting intents present in heterogeneous datasets.

Our contribution explicitly builds upon this modern architectural foundation. We extend this line of work by shifting IQ-Learn's inherently zero-centered penalty into an informed prior centered on an ego-centric social scoring function. By using this social impact signal as an informed Tikhonov prior, we combine social preference specification with principled regularization of the reward recovery process.

4

Methodology

This chapter presents the proposed Social Impact Regularized IQ-Learn framework. The primary objective of this methodology is to enable the recovery of specific, socially aligned driving behaviours from heterogeneous demonstration datasets.

To achieve this, the framework integrates a normative prior directly into the inverse reinforcement learning process. The high level architecture of this approach is illustrated in Figure 4.1. As shown in the diagram, the pipeline explicitly differentiates between the raw data inputs and the active algorithmic steps. It begins with the ingestion of mixed expert trajectories, which serve purely as the foundational data input. Rather than attempting to fit a single reward function to these inherently conflicting behaviours, the framework first applies a reward decomposition, separating the underlying objectives into an individual driving component and an ego-centric social impact component. These decomposed signals are then combined to form a social scoring function parameterized by a designated social target. This scoring function acts as the normative prior. Finally, this prior is integrated into the Regularized IQ-Learn optimization process, which balances the extraction of general driving dynamics from the full dataset with the enforcement of the target social profile, ultimately yielding a socially aligned reward function as the final system output.

The remainder of this chapter is structured to follow this pipeline. Section 4.1 defines the social reward decomposition. Then Section 4.2 explains the scoring function used. Finally, Section 4.3 presents the algorithmic contribution, detailing the mathematical formulation of the reward regularization and its integration into the IQ-Learn objective.

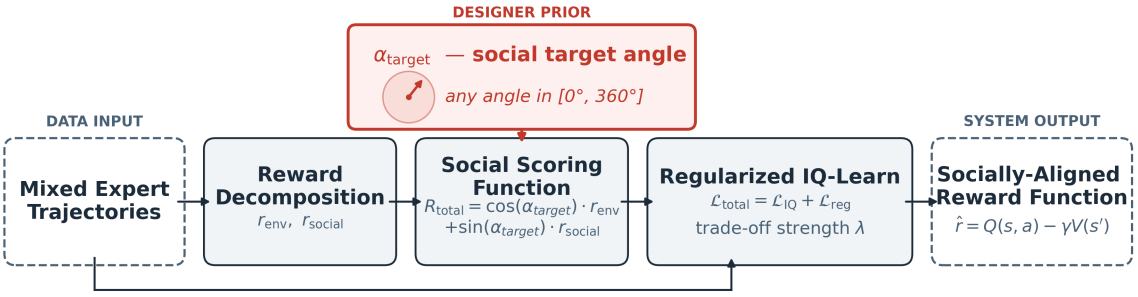


Figure 4.1: Overview of the proposed Social Impact Regularized IQ-Learn framework. Mixed expert demonstrations are decomposed into individual and social reward components, combined into a social scoring function, and incorporated as a regularization prior within the IQ-Learn optimization process to recover a socially aligned reward function.

4.1. Reward Decomposition

Each transition in the demonstration dataset is annotated with two reward signals, both computed from the ego vehicle's perspective.

The individual reward $r_{env}(s,a)$ captures the ego vehicle's standard driving objectives. It is a composite signal consisting of a progress component that rewards longitudinal progress along the ego's planned route, a speed component that encourages maintaining a target velocity, a collision penalty, a constant per-step penalty to discourage unnecessary idling and an arrival bonus upon reaching the destination. Formally,

$$r_{env}(s,a) = (w_p \cdot \Delta_{progress}(s) + w_v \cdot \bar{v}(s)) \cdot \mathbb{I}_{on\_road}(s) + w_s, \tag{4.1}$$

where $\Delta_{progress}$ denotes the normalized longitudinal advance along the ego's route, $\bar{v}$ is the speed clipped and scaled to $[0,1]$, $\mathbb{I}_{on\_road}$ is 1 when the vehicle remains on the road and 0 otherwise, and $w_s < 0$ is the per-step cost. Collision and arrival are handled as terminal overrides: if the ego collides $r_{env} = w_c$. If it arrives at the destination, $r_{env} = w_a$. Finally if the episode times out without either event, an additional penalty w_t is applied.

The social impact reward $r_{social}(s,a)$ measures the effect of the ego vehicle's actions on its neighbours. We define the influence zone as the region of the state space where the ego agent is positioned to meaningfully affect neighbour behaviour. This captures states of approaching or ongoing interactions. In the intersection environments used in our experiments, this signal is based on counterfactually inspired yielding - it is nonzero only when the ego is in the influence zone, approaching or occupying the intersection conflict area while there is traffic nearby.

When the ego decelerates or stops while neighbours are approaching the conflict zone, it is yielding, sacrificing its own speed to allow others to proceed safely. This produces a positive r_{social}. Conversely, when the ego maintains speed and forces neighbours to brake, it is causing a disruption and the ego's actions impose a cost on others. This produces a negative r_{social}. When the ego is far from any potential interaction $r_{social} = 0$. This design ensures that social reward cannot be accumulated passively. It can only be earned or lost through active interaction with other traffic participants. Formally the social impact reward is

$$r_{social}(s,a) = w_y \cdot (1 - v_{ratio}) \cdot n + w_d \cdot \sum_{i \in N} \min\left(\frac{\max(\Delta v_i, v_{\text{thresh}})}{v_{\text{cap}}}, 1\right), \tag{4.2}$$

where $v_{ratio} = v_{ego}/v_{max}$ is the ego's speed relative to the maximum, n is the number of neighbours within the conflict distance, $\Delta v_i = v_i^{\text{prev}} - v_i^{\text{curr}}$ is the speed loss of neighbour i between consecutive timesteps, $v_{\text{thresh}} = 1.0\,\text{m/s}$ is a dead-zone threshold below which disruption is not counted, $v_{\text{cap}} = 5.0\,\text{m/s}$ clips the maximum attributable speed loss, and $w_y > 0, w_d < 0$ are the yielding reward and disruption penalty weights, respectively. The yielding term scales linearly with the number of nearby neighbours, reflecting that yielding is more socially significant when more vehicles benefit. The disruption term activates only when a neighbour's speed loss exceeds the threshold, preventing normal traffic fluctuations from being attributed to the ego. Mirroring the structure of r_{env} any neighbour collision caused by the ego agent incurs a severe terminal penalty.

4.2. The Social Scoring Function

Given the two reward components, we define a social scoring function for each transition using the geometric projection

$$R_{\text{total}}(s, a \mid \alpha) = \cos(\alpha) \cdot r_{\text{env}}(s, a) + \sin(\alpha) \cdot r_{\text{social}}(s, a), \tag{4.3}$$

where $\alpha \in [0, 2\pi)$ is the target social angle specified by the system designer. This scoring function is not used to generate behaviour. It serves exclusively as a regularization target during reward learning: it encodes the designer's preference for how strongly the recovered reward should reflect the ego's social impact relative to its individual objectives.

The choice of geometric projection is motivated by two properties. First, the cosine-sine weighting provides a smooth rotation between purely individual ($\alpha = 0°$, where $R_{\text{total}} = r_{\text{env}}$) and purely social ($\alpha = 90°$, where $R_{\text{total}} = r_{\text{social}}$) objectives, with all intermediate trade-offs accessible by varying α. Second, the angle is recoverable from observed behaviour via $\hat{\alpha} = \text{atan2}(\bar{R}_{\text{social}}, \bar{R}_{\text{env}})$, which provides a natural verification metric for whether the regularization achieved the intended social alignment.

4.3. Regularized IQ-Learn

With the social scoring function R_{total} defined, the design question becomes where in the IQ-Learn learning process should the social prior enter? We propose integrating the social prior via Reward Regularization, which modifies the IQ-Learn objective $\mathcal{L}_{IQ}$ (Equation 2.7). Alternative strategies, such as Bellman target shifting and demonstration reweighting were also explored but present distinct mathematical and practical limitations. These are detailed in Appendix B.

4.3.1. Reward Regularization

This integration strategy applies Tikhonov regularization [24] with an informed prior to the IQ-Learn reward recovery. As introduced in Section 2, IQ-Learn's χ^2 term already implements Tikhonov regularization with a zero-centered prior, penalizing reward functions of large magnitude. Our contribution is to supplement this with an informed prior centered on the social scoring function, expressing the system designer's preference for reward functions that reflect a specific social alignment.

The choice of a Tikhonov (quadratic) penalty rather than alternative regularizers is deliberate. An L_1 penalty $\|\hat{r} - \lambda R_{\text{total}}\|_1$ would induce sparsity in the deviation from the prior, encouraging the recovered reward to match R_{total} exactly on most state-action pairs and depart sharply on others. This is a feature-selection bias and is inconsistent with our modelling assumption that the social target acts as a soft preference rather than a hard constraint. Information theoretic penalties such as the Kullback-Leibler divergence are not directly applicable, since $\hat{r}$ and R_{total} are reward functions rather than probability distributions and would require an additional softmax normalization that distorts the semantic alignment we wish to enforce. Alternative f-divergences in the spirit of IQ-Learn's χ^2 term (Section 2.2.2) could in principle be shifted to a non-zero prior, but they share the same Bayesian interpretation as the quadratic penalty only in the Gaussian limit, while introducing additional hyperparameters and asymmetries. Consequently, the quadratic penalty uniquely satisfies the demands of our framework by providing an exact Gaussian prior, maintaining architectural continuity with IQ-Learn's native χ^2 regularizer,

ensuring smooth optimization dynamics, and conferring classical well-posedness guarantees to the inverse problem.

Concretely, we add an auxiliary mean squared error loss that operates directly on the implicit reward recovered by IQ-Learn, pushing it toward the social scoring function:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\rho_E}\left[\left(\hat{r}(s,a) - \lambda \cdot R_{\text{total}}(s,a \mid \alpha_{\text{target}})\right)^2\right]. \tag{4.4}$$

The total training objective becomes:

$$\mathcal{L}_{total} = \mathcal{L}_{IQ} + \mathcal{L}_{reg}. \tag{4.5}$$

Beyond its mathematical foundation, this formulation has four desirable properties that distinguish it from the alternatives. First, the social regularization is additive to the IQ-Learn loss. The IQ-Learn objective continues to fit the expert data, ensuring that the recovered reward captures the physical driving dynamics present in all demonstrations, while the MSE provides an independent gradient signal about the social character of the reward. These two pressures can be balanced via λ without one corrupting the internal consistency of the other. Crucially, because $\mathcal{L}_{reg}$ is a separate term, the χ^2 regularization in Term 3 continues to operate on the unshifted implicit reward $\hat{r}$ exactly as in standard IQ-Learn, preserving the theoretical bound on learned reward magnitude established by [22]. Second, unlike reweighting, every transition in the expert buffer contributes equally to $\mathcal{L}_{IQ}$. Physical driving competence is learned from the full dataset regardless of the social alignment of the demonstrator. The MSE loss then selects, among the set of reward functions that adequately explain the data, the one whose implicit reward is closest to the desired social profile. Third, the gradient of the MSE loss with respect to the Q-network parameters θ is:

$$\frac{\partial \mathcal{L}_{\text{reg}}}{\partial \theta} = 2\mathbb{E}_{\rho_E}\left[(\hat{r}(s,a) - \lambda \cdot R_{\text{total}}(s,a)) \cdot \frac{\partial \hat{r}(s,a)}{\partial \theta}\right], \tag{4.6}$$

which directly pushes the Q-network toward producing an implicit reward that matches R_{total} on expert states. This is a far stronger learning signal than the indirect effect of observing certain transitions more frequently, because it acts on every update step and targets the reward structure explicitly. Finally, the regularization strength λ controls the trade-off. At $\lambda = 0$, we recover standard IQ-Learn, but as λ increases, the recovered reward is increasingly constrained to match the social target. This, however, comes at the potential cost of reduced fidelity to the demonstration data.

5

Experimental Setup

This chapter details the experimental setup used to evaluate the social impact regularized IQ-Learn framework. A key consideration in this evaluation is ensuring that the explicit effects of the social regularizer are not obscured by the complexities of deep function approximation. Evaluating the social regularizer exclusively in a high dimensional continuous simulator would make it nearly impossible to decouple the mathematical efficacy of the algorithm from neural network optimization dynamics or environmental noise. To rigorously isolate these factors, our evaluation uses a deliberate progression across three environments. This structure acts as a methodological baseline: it first mathematically verifies the regularizer's mechanics under exact, tabular conditions, before progressively introducing multi-agent stochasticity and finally testing its scalability with continuous kinematics.

First, a preliminary 5×5 deterministic gridworld serves as a proof-of-concept. Its fully enumerable state space allows for the exact computation of optimal expert policies via value iteration. This provides a mathematically perfect ground truth baseline, allowing us to verify the regularizer's core mechanics and the agent's behavioural shift before introducing the confounding complexities of high dimensional continuous observations and naturalistic traffic dynamics. Second, a 7×7 gridworld introduces multi-agent stochasticity, yielding dynamics, and discrete speed choices. This bridges the gap to interactive traffic scenarios while remaining analytically tractable. Finally, we evaluate the framework in the continuous observation Intersection-v1 scenario from Highway-Env [37]. This high dimensional environment tests the algorithm's scalability to neural networks and naturalistic traffic behaviours.

The following sections outline these environments, the expert data generation process, dataset construction, training configurations, and the metrics used to evaluate both driving competence and social alignment.

5.1. Environments

5.1.1. Gridworld Proof-of-Concept

Our preliminary proof-of-concept environment is a deterministic 5×5 gridworld that models a minimal unsignalized intersection, shown in Figure 5.1. The ego vehicle begins at position $(0, 2)$ travelling East, while a single reactive opponent vehicle begins at $(2, 0)$ travelling North. Both trajectories converge at a shared conflict point in cell $(2, 2)$. At each timestep, the ego chooses between two discrete actions: *Wait* (remain stationary) or *Forward* (advance one cell). Meanwhile, the other vehicle follows a fixed forward policy subject to a simple reactive rule: it holds its position for that step if advancing would cause simultaneous occupancy of the conflict cell. Episodes are capped at 20 timesteps. The observation space is a 4-dimensional vector $(x_{\text{ego}}, y_{\text{ego}}, x_{\text{other}}, y_{\text{other}})$. The reward structure is decomposed into two independent components: r_{env} grants +10 upon the ego's arrival and a -1 penalty per step, while r_{social} yields +10 upon the other vehicle's arrival and a -1 penalty per step. The influence zone covers the conflict cell $(2, 2)$ and the cell immediately preceding it along the ego's path, becoming active only when

the opponent vehicle is within reach of the conflict point.

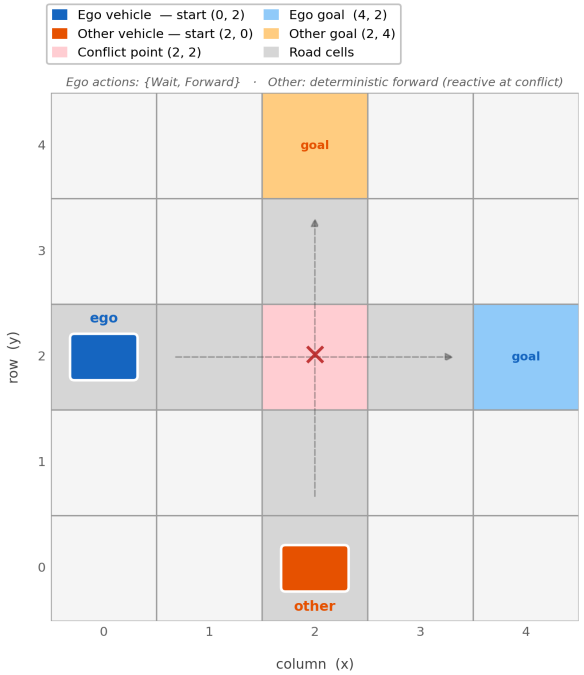


Figure 5.1: Deterministic 5×5 gridworld proof-of-concept. This environment models a minimal unsignalized intersection where an ego vehicle travelling East and a single reactive opponent travelling North converge at a shared conflict point.

The simplicity of this environment ensures analytical tractability. Optimal policies for any target angle α_{target} can be computed exactly via value iteration over the small, fully enumerable state space. This provides ground-truth behaviour to cleanly verify the regularizer's effect and serve as a proof-of-concept.

5.1.2. 7x7 Gridworld

The first experimental environment is a 7×7 gridworld modelling an unsignalized intersection with reactive cross-traffic, shown in Figure 5.2. This environment is designed to bridge the gap between a minimal proof-of-concept and a continuous driving simulator. It introduces multi-agent stochasticity, speed-based decisions, and an ego-centric social reward structure, while still remaining tractable for exact solution via value iteration.

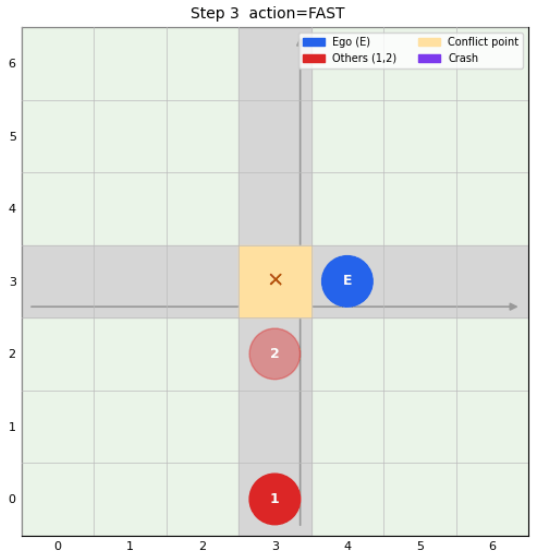


Figure 5.2: 7×7 multi-agent stochastic gridworld. This environment models an unsignalized cross-shaped intersection. An ego vehicle travels West to East along row 3, while higher-priority, reactive cross-traffic moves South to North along column 3. Both paths meet at a shared conflict point located at cell (3, 3).

The grid consists of two perpendicular roads forming a cross-shaped intersection. The ego vehicle travels West to East along row 3, while other vehicles travel South to North along column 3, with the conflict point at cell (3, 3). The ego has lower priority. Other vehicles have right of way, and they react to the ego's presence only when a collision is about to happen. The action space consists of three discrete speed choices: STOP (0 cells/step), SLOW (1 cell/step), and FAST (2 cells/step), mirroring the three target speeds used in the intersection environment. Two other vehicle slots are available: the first always spawns at the southern edge, while the second spawns stochastically with probability $p = 0.3$ per timestep when the slot is empty. Other vehicles follow IDM-like reactive behaviour: they maintain their desired speed (FAST) but decelerate when the ego approaches the conflict zone, mimicking the yielding dynamics of real traffic. Episodes terminate when the ego reaches the eastern edge (arrival), occupies the conflict point simultaneously with another vehicle (collision), or exceeds 25 timesteps (timeout).

The observation is a 6-dimensional vector $(x_{\text{ego}}, v_{\text{ego}}, y_1, v_1, y_2, v_2)$, where x_{ego} and v_{ego} are the ego's position and speed, and (y_i, v_i) are the position and speed of each other vehicle (with $y_i = -1$ indicating that the vehicle has not yet spawned). With 7 positions, 3 speeds, and 8 possible position values per other vehicle (including the not-spawned), the state space contains 12,096 states — small enough to be tractable for exact value iteration, yet large enough to exhibit nontrivial multi-agent dynamics with stochastic spawning.

The ego's individual reward r_{env} consists of a speed component ($w = 0.4 \times v/v_{\max}$), a collision penalty (-5.0), an arrival bonus (+2.0), and a per-step cost (-0.3). The social impact reward r_{social} follows the reaction-based design described in Section 4.1: a yielding reward ($+0.2 \times n_{\text{approaching}}$) when the ego stops or slows near the conflict zone while others approach, a disruption penalty (-1.0 per speed level of forced deceleration, -1.5 for a complete stop), and a neighbour collision penalty (-5.0). These weights were chosen to ensure social signal separability, as verified in Appendix A. For this environment, the influence zone consists of the conflict cell (3, 3) and the cell immediately preceding it along the ego's path.

5.1.3. Highway-Env: Intersection-v1

The third experimental environment is the intersection-v1 scenario from the Highway-Env simulator [37]. This environment tests whether the regularization framework transfers from tabular settings to

function approximation with neural Q-networks on high-dimensional observations.

Intersection-v1 provides a continuous kinematic simulation of an unsignalized four-way intersection, as shown in Figure 5.3. Vehicles follow lane-based trajectories with realistic acceleration and steering dynamics. Cross-traffic is controlled by IDM-based drivers with randomized parameters. Each spawned vehicle samples its desired speed, time headway, jam distance, acceleration capability, and comfortable deceleration from calibrated distributions based on [38], introducing naturalistic behavioural diversity into the traffic.

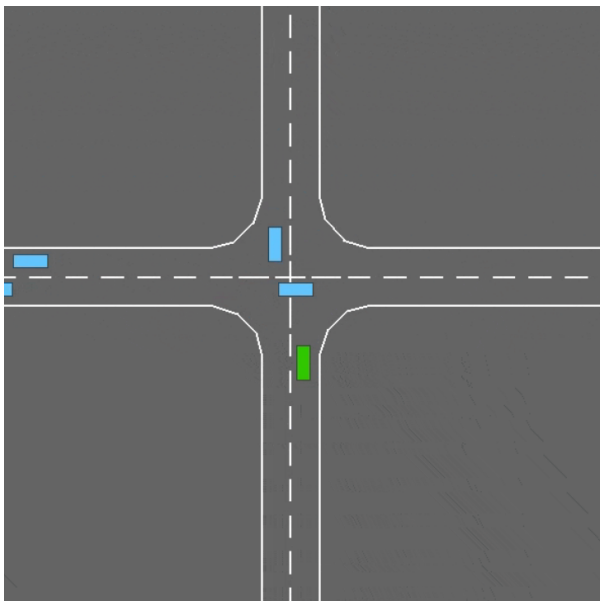


Figure 5.3: Intersection-v1 continuous simulation environment. This environment provides a continuous kinematic simulation of an unsignalized four-way intersection. The layout features lane-based vehicle movement with realistic acceleration and steering dynamics, where the ego vehicle navigates through cross-traffic controlled by randomized IDM-based drivers.

The ego vehicle observes the kinematics of up to 10 vehicles (including itself) using a 7-feature representation per vehicle: presence flag, absolute position (x, y), velocity (v_x, v_y), and heading $(\cos h, \sin h)$. The observation is flattened into a 70-dimensional vector serving as input to the Q-network.

The ego selects from three discrete longitudinal actions corresponding to target speeds of 0, 4.5, and 9 m/s. Lateral steering is handled automatically by the lane-following controller. The simulation runs at 10 Hz with policy decisions at 1 Hz, and episodes last up to 20 seconds.

Each episode starts with 5 vehicles, and additional vehicles spawn with probability 0.3 per timestep. The ego's destination is fixed, and the environment uses a single controlled vehicle.

The environment's built-in reward is replaced entirely by a custom wrapper that computes the reward decomposition described in Section 4.1. The individual reward r_{env} uses a progress-based structure with weights $w_p = 2.0$ (progress), $w_v = 0.5$ (speed), $w_c = -2.0$ (collision), $w_a = 1.0$ (arrival), $w_s = -0.15$ (per-step cost), and $w_t = -0.5$ (timeout). The influence zone in this environment covers the final portion of the ego's approach lane (the segment beyond 8 m of longitudinal travel) and the internal lane segments traversing the intersection, with cross-traffic counted as relevant when within 20 m of the ego. The social impact reward r_{social} uses the same mechanism and weights as the gridworld.

5.2. Expert Policy Generation and Dataset Construction

The demonstrations dataset is generated by sampling trajectories from synthetic expert policies. For each environment, we train separate experts at three distinct social angles: egoistic ($\alpha = 0°$), prosocial ($\alpha = 45°$), and altruistic ($\alpha = 90°$). Synthetic experts are used instead of naturalistic human driving data to provide mathematical ground truths for the underlying social intent. This enables exact quantification of how well the regularizer recovers the target behavioural mode, isolating the algorithm's performance from the noise of subjective human dataset labelling.

In the gridworld environments, optimal policies are computed via exact value iteration. Because the state space is fully enumerable, the optimal value function is first computed on a deterministic variant of the environment, sweeping until convergence. Demonstrations are collected by rolling out these optimal policies in the stochastic variant. This ensures the expert exhibits exact theoretical alignment with its assigned social angle while still encountering environment induced variance.

For the continuous-observation Intersection-v1 environment, expert policies are trained using Deep Q-Networks (DQN). DQN was selected because the environment relies on discrete longitudinal actions, and a standard value-based method provides a stable expert baseline that conceptually mirrors the subsequent IQ-Learn architecture used by the learner. The experts utilize a two-layer multi-layer perceptron (256 units each), a learning rate of 3×10^{-4}, and a discount factor of $\gamma = 0.99$. Training spans 5,000,000 timesteps distributed across 12 parallel environments, utilizing the combined R_{total} as the standard RL reward signal.

Upon completing expert training, policy rollouts are gathered to construct the demonstration dataset. Each transition is recorded as an extended tuple: $(s, a, r, s', d, r_{env}, r_{social}, G_{env}, G_{social})$. To facilitate the regularization framework, the individual reward components (r_{env} and r_{social}) and their cumulative returns are preserved independently, rather than being collapsed into a single scalar R_{total}. This decoupled structure allows the regularizer to dynamically compute the social scoring function for any target angle during the IRL training phase. In this way, the framework avoids the need for repeated data collection, improving both computational efficiency and experimental flexibility.

Finally, to emulate the heterogeneity of real-world driving data, these trajectories are aggregated into a mixed dataset. The primary configuration is a uniform mixture containing equal proportions of egoistic, prosocial, and altruistic demonstrations. This balanced distribution ensures that no single behavioural mode statistically dominates the training data. This allows us to establish a rigorous baseline for evaluating the regularizer: if the framework successfully isolates and recovers a specific social profile from this mixture, it confirms that the mechanism actively steers the policy based on the social prior, rather than passively exploiting imbalances in the data.

5.3. IQ-Learn Configuration

The IQ-Learn agent uses a DQN backbone with the soft value function $V(s) = \tau \log \sum_a \exp\left(\frac{Q(s,a)}{\tau}\right)$ and a soft updated target network (Polyak averaging). All experiments use the χ^2 divergence and the 'value' sampling strategy following [22]. The social regularization mode used is 'reward_reg' with cumulative returns. λ is the primary experimental variable and is swept across experiments. Appendix C summarizes all hyperparameters.

5.3.1. Practical Considerations

The implicit reward $\hat{r} = Q(s, a) - \gamma(1 - d)V(s')$ is an estimate of instantaneous reward, but because Q-values represent cumulative discounted returns, $\hat{r}$ inherits a scale that reflects trajectory level quantities. When the social scoring function is computed from instantaneous per-transition values (r_{env}, r_{social}), a significant scale mismatch can arise between $\hat{r}$ and R_{total}, potentially requiring explicit normalization. However, when computed from cumulative discounted returns $R_{total} = \cos(\alpha) \cdot G_{env} + \sin(\alpha) \cdot G_{social}$, the scoring function naturally operates on a scale similar to the Q-values. The MSE does not assert

that the true reward is R_{total}, but rather expresses a preference among the set of reward functions that explain the data, and combining cumulative returns with the scaling parameter λ naturally aligns the regularization target with the implicit reward, eliminating the need for explicit normalization.

5.4. Evaluation Metrics

To verify that the regularization achieves the intended social alignment without degrading driving competence, the learned policies are evaluated across a standardized set of core metrics applicable to all environments.

First, social alignment is quantified using the empirical angle. This is computed from the policy's mean accumulated rewards over the evaluation episodes:

$$\hat{\alpha} = \text{atan2}(\bar{R}_{\text{social}}, \bar{R}_{\text{env}})$$

Successful regularization results in $\hat{\alpha} \approx \alpha_{\text{target}}$. Significant deviations from this target indicate either under-regularization, where the agent converges to an arbitrary behavioural mode, or over-regularization, where the policy collapses into degenerate behaviour.

Second, fundamental physical driving competence is measured by tracking the ego vehicle's collision rate, arrival rate, and timeout rate. Additionally, progression metrics such as mean speed, percentage of time stopped, and overall episode length are recorded to ensure physical driving abilities are actively preserved rather than sacrificed for extreme caution. To account for traffic stochasticity, all results are aggregated across 5 random seeds. The primary baseline for comparison is the unregularized IQ-Learn agent, with the original expert policies provided as boundary references.

6

Results

This chapter presents the empirical evaluation of the social impact regularized IQ-Learn framework. The primary objective of these experiments is to demonstrate that the proposed regularizer successfully steers the recovered policy toward a targeted social alignment and to quantify the resulting trade-offs between physical driving competence and social behaviour.

To provide a clear analysis of both the algorithm's mechanics and its scalability, the results are structured across the progressive evaluation pipeline established in the previous chapter. First, we conduct a preliminary validation on a minimal deterministic gridworld to verify the regularizer's core mechanism and its basic ability to alter learned behaviour. Second, we expand the evaluation to a multi-agent stochastic gridworld, verifying its ability to isolate distinct behavioural modes and examining its sensitivity to the regularization strength. Finally, we extend the evaluation to the high-dimensional Intersection-v1 simulator to confirm that these tabular findings successfully transfer to a continuous kinematic environment relying on deep function approximation. Across all domains, the empirical findings confirm that the regularizer effectively resolves the behavioural ambiguity inherent in heterogeneous driving datasets.

6.1. Proof-of-Concept Demonstration

To verify that the MSE regularizer can alter learned behaviour at a basic level, the framework was first tested on a 5×5 deterministic gridworld. This environment features a single ego vehicle, a single reactive cross-traffic vehicle, and a shared conflict point. Optimal policies for different target angles were computed via value iteration, and demonstrations were extracted for IQ-Learn training. The resulting action sequences (Figure 6.1) confirm that the regularizer successfully shifts the agent's decisions based on the designated target angle.

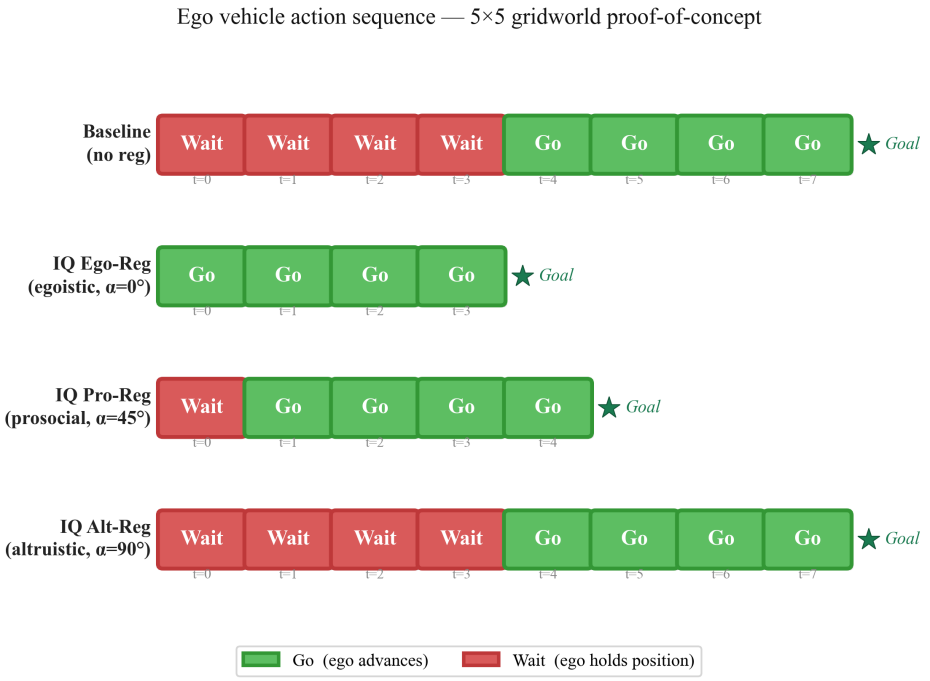


Figure 6.1: Preliminary validation of the Social Impact Regularized IQ-Learn framework in the 5×5 proof-of-concept gridworld. Policies learned from the same mixed demonstration dataset exhibit distinct behavioural modes as the target social angle increases from egoistic to altruistic, demonstrating the regularizer's ability to steer social alignment.

When trained on the mixed dataset without the social impact regularizer, the IQ-Learn baseline converges to a highly cautious yielding policy. The observed action sequence (Figure 6.1, topmost panel) shows the baseline agent holding its position at the start by executing four consecutive "Wait" actions. It only executes "Go" actions to advance toward its goal after the crossing vehicle has fully cleared the intersection.

Applying the reward regularization framework with three distinct target angles (α_{target}) to the same mixed dataset produces strictly separated behavioural modes.

First, we evaluate the egoistic target profile ($\alpha_{target} = 0°$). The regularized agent adopts an assertive driving profile, executing continuous "Go" actions from the first timestep without yielding. Because the cross-traffic vehicle is programmed to hold its position if advancing would cause a collision, the egoistic agent forces the other vehicle to wait while it clears the conflict point first.

Second, we evaluate the prosocial target profile ($\alpha_{target} = 45°$). The recovered policy demonstrates an intermediate, selective yielding behaviour. The agent executes a single "Wait" action at the beginning of the episode, allowing the other vehicle to safely pass through the conflict point, before proceeding with consecutive "Go" actions to complete its own trajectory.

Finally, we evaluate the altruistic target profile ($\alpha_{target} = 90°$). The regularized agent exhibits caution, adopting an action sequence that is identical to the unregularized baseline policy. It holds its position with four consecutive "Wait" actions to ensure zero disruption to the cross-traffic before eventually advancing through the intersection.

6.2. 7x7 Gridworld

Following the initial demonstration, the evaluation is extended to the 7×7 gridworld environment. As established in Section 5.1.2, this scenario introduces multi-agent stochasticity, discrete speed choices,

and reactive cross traffic dynamics. This serves as an intermediate step that bridges the complexity gap toward a continuous simulator while remaining an analytically tractable, discrete environment. The following analysis is divided into two parts: first, establishing the baseline behaviour of the unregularized framework on mixed data, and second, evaluating the regularizer's capacity to isolate and steer policies toward distinct target angles from the exact same mixed data.

We begin by examining what unregularized IQ-Learn produces when trained on the mixed dataset containing equal parts egoistic, prosocial and altruistic demonstrations. The baseline agent recovers an empirical angle of $\hat{\alpha} \approx 152°$ (Figure 6.2, grey square), placing it in the second quadrant of the social ring. Looking at the action heatmap (Figure 6.4, leftmost panel), this manifests as a policy that always yields. The agent stops before the intersection regardless of whether traffic is present or not, and never crosses.

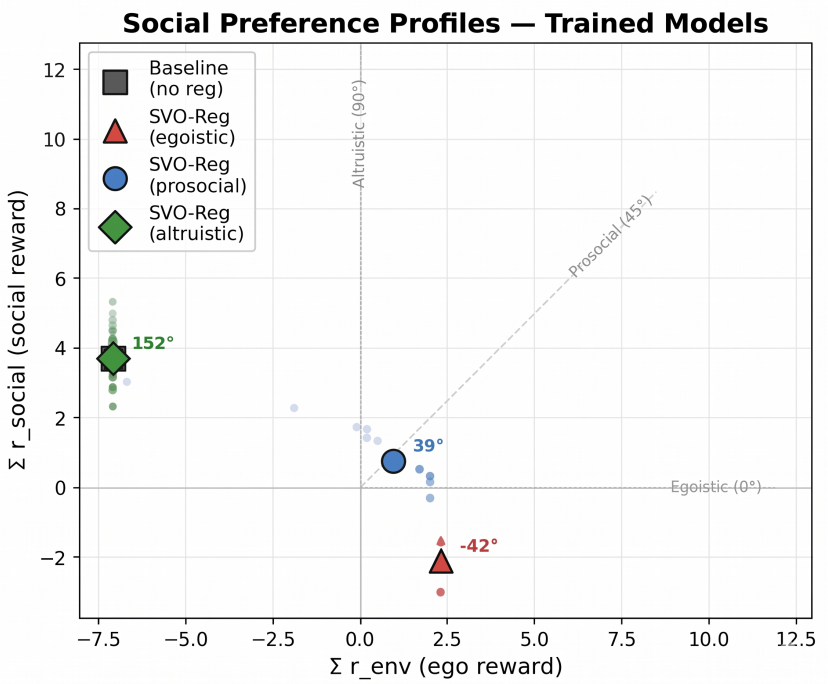


Figure 6.2: Social preference profiles of the trained models. This plot illustrates the empirical angles recovered by the baseline and regularized IQ-Learn agents when trained on the mixed dataset. It maps the positions of the unregularized baseline agent (grey square) alongside the agents steered toward egoistic (red triangle), prosocial (blue circle), and altruistic (green diamond) targets.

We now apply the reward regularization with three different target angles, each trained on the same mixed dataset to test whether the regularizer can recover distinct behavioural modes. As a point of reference, the ground truth behaviours for each pure mode are established via Value Iteration (Figure 6.3).

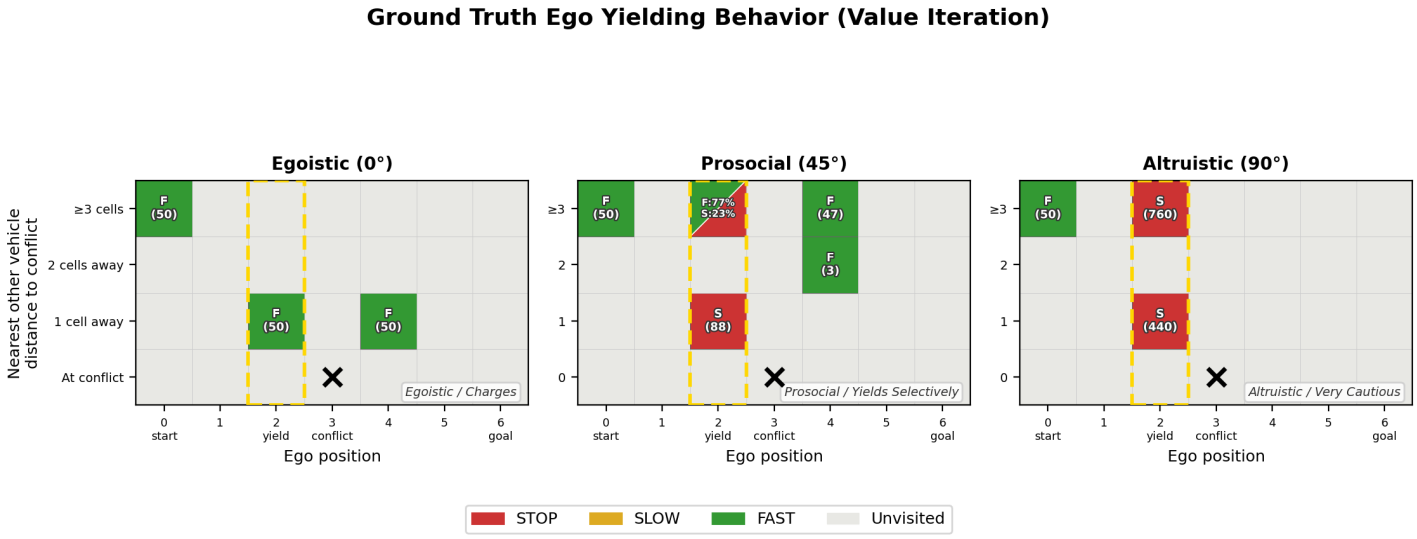


Figure 6.3: Ground truth behaviour. This figure illustrates the optimal policy state-action mappings derived via value iteration for the three pure social profiles: egoistic, prosocial, and altruistic.

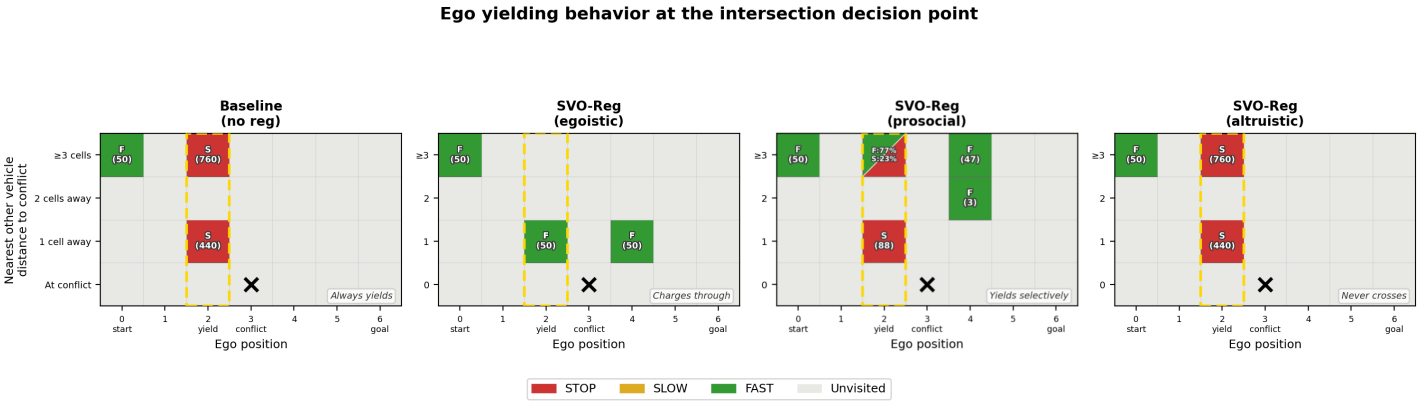


Figure 6.4: Learned action policies of the trained agents. This figure displays the action decisions executed by the unregularized baseline model alongside the three regularized models (egoistic, prosocial, and altruistic) trained on the mixed dataset.

First, we start with the egoistic target ($\alpha_{target} = 0°$). The regularized agent recovers an empirical angle of $\hat{\alpha} \approx -42°$ (Figure 6.2, red triangle). The action heatmap (Figure 6.4, second panel) confirms that the agent rushes through the intersection at full speed, regardless of cross traffic. It never yields.

Second, we move on to the prosocial target ($\alpha_{target} = 45°$). The regularized agent recovers $\hat{\alpha} \approx 39°$ (Figure 6.2, blue circle). The action heatmap (Figure 6.4, third panel) reveals the most nuanced behaviour - the agent yields selectively. When cross traffic is close (1 cell away), the agent stops. When traffic is more distant (≥3 cells), the agent shows mixed behaviour, sometimes crossing and sometimes stopping.

Finally, we take a look at the altruistic target ($\alpha_{target} = 90°$). The regularized agent recovers $\hat{\alpha} \approx 152°$ (Figure 6.2, green diamond), which is exactly equal to the baseline. The action heatmap (Figure 6.4, rightmost panel) shows identical behaviour with the baseline, the agent stops before the intersection and never crosses.

To understand how the regularization strength λ controls the transition from baseline to target behaviour, we conduct a sweep over different λ values on the egoistic target angle. The empirical results of this

sweep, detailing the behavioural shifts across regularization strengths are reported in Appendix D.

6.3. Intersection-v1

Following the discrete multi-agent evaluation in the 7×7 gridworld, the final stage of the evaluation is conducted in the Intersection-v1 environment. As detailed in Section 5.1.3, this scenario provides a continuous kinematic simulation with high dimensional observations, testing the framework's scalability to deep neural network function approximation and naturalistic traffic dynamics. The following subsections first analyse the baseline IQ-Learn performance when trained on mixed demonstration data. Subsequently, the regularized agents are compared against the original expert policies to evaluate both the preservation of physical driving competence and the successful enforcement of the designated social profile.

The unregularized IQ-Learn baseline, trained on the mixed intersection dataset, again consisting of equal parts egoistic, prosocial and altruistic transitions, produces a policy with a collision rate of 18.0% (± 2.6%), an arrival rate of 23.6% (± 1.4%), and a timeout rate of 58.4% (± 2.0%) across five evaluation seeds (Table 6.1). The agent drives at a moderate average speed of 10.2 km/h and spends 51.5% of its time stopped. Notably, the recovered empirical α angle collapses towards the prosocial behaviour in this environment (Figure 6.5, grey square).

Metric	Mean	± Std (across seeds)
Collision Rate	18.0%	± 2.6%
Arrival Rate	23.6%	± 1.4%
Timeout Rate	58.4%	± 2.0%
Avg Speed	10.2 km/h	± 0.4 km/h
Time Stopped (%)	51.5%	± 1.0%

Table 6.1: Aggregated baseline results (IQ-Learn) for the intersection environment across 5 seeds.

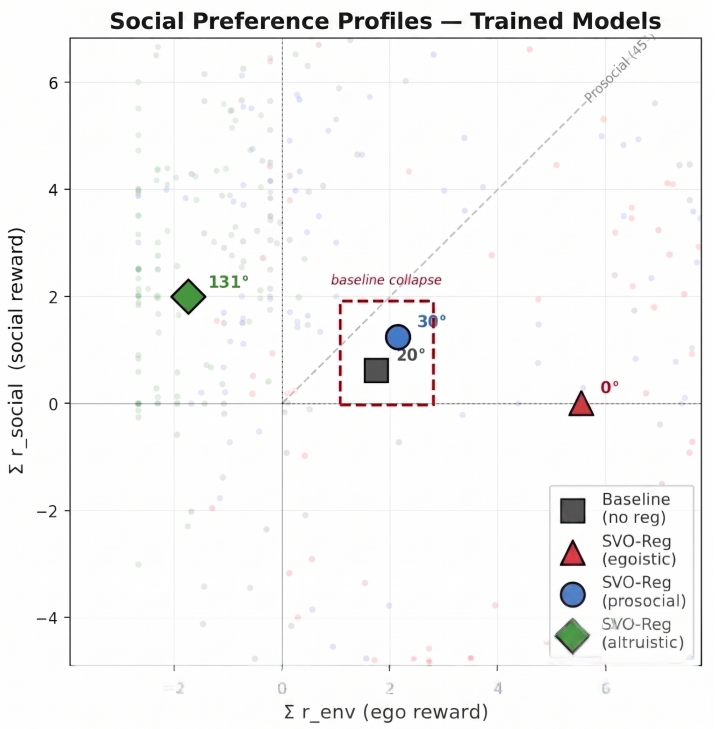


Figure 6.5: Social preference profiles. This plot illustrates the empirical angles recovered by the baseline and regularized IQ-Learn agents when trained on the mixed Intersection-v1 dataset. It maps the positions of the unregularized baseline agent (grey square) alongside the agents steered toward egoistic (red triangle), prosocial (blue circle), and altruistic (green diamond) targets.

Now we take a look at the regularized agents. Table 6.2 compares each regularized IQ-Learn agent against its corresponding expert policy. All agents are trained on the same mixed dataset and differ only in the target angle α_{target}.

First, we take a look at the egoistic regularization ($\alpha_{\text{target}} = 0°$). The IQ Ego-Reg agent matches the expert's collision rate (29% vs 29%) but with a lower arrival rate (55% vs 70%) and a higher timeout rate (16% vs 1%). It drives at 20 km/h compared to the expert's 28 km/h, and spends 21% of its time stopped versus the expert's 2%. The recovered empirical angle is 0°, matching perfectly the target angle (Figure 6.5, red triangle).

The prosocially regularized agent ($\alpha_{\text{target}} = 45°$) achieves a collision rate of 23±2.5%, which is lower than the prosocial expert's 25±1.5%. The arrival rate is reduced (46% vs 54%) and the timeout rate increased (32% vs 21%), indicating that the regularized agent is somewhat more cautious. Both drive at 16 km/h with 33% time stopped. The recovered empirical angle is 30°, staying close to the 45° target angle (Figure 6.5, blue circle).

Finally, the altruistically regularized ($\alpha_{\text{target}} = 90°$) agent achieves the lowest collision rate across all agents (11%), dramatically improving on the altruistic expert's 29%. However, this comes at a severe cost to driving competence: the arrival rate is only 1%, the timeout rate is 88%, speed is 3.8 km/h, and the agent spends 71% of its time stopped. The agent has effectively adopted stopping as its primary strategy. The recovered empirical angle, therefore, is 131°, being a bit off the target 90° angle (Figure 6.5, green diamond).

Table 6.2: Performance metrics comparison across expert baselines and regularized IQ-Learn models.

Metric	Expert Ego	IQ Ego-Reg	Expert Pro	IQ Pro-Reg	Expert Alt	IQ Alt-Reg	IQ Base
Collision	29 ± 5.4%	29 ± 4.0%	25 ± 1.5%	**23 ± 2.5**%	29 ± 3.1%	**11 ± 1.7**%	18.0 ± 2.6%
Arrival	70 ± 5.3%	55 ± 3.3%	54 ± 2.7%	46 ± 1.7%	0 ± 0.0%	**1 ± 1.2**%	23.6 ± 1.4%
Timeout	1 ± 0.4%	16 ± 2.3%	21 ± 3.9%	32 ± 1.7%	71 ± 3.1%	88 ± 1.9%	58.4 ± 2.0%
Speed	28 ± 0.3 km/h	20 ± 0.6 km/h	16 ± 0.2 km/h	16 ± 0.3 km/h	5.8 ± 0.1 km/h	3.8 ± 0.3 km/h	10.2 ± 0.4 km/h
Stopped	2 ± 0.3%	21 ± 1.6%	33 ± 0.6%	33 ± 1.3%	62 ± 1.0%	71 ± 1.8%	51.5 ± 1.0%

7

Discussion

This section interprets the experimental results in the context of the research questions posed in the introduction, identifies the practical implications of the findings and connects them to the claims made earlier.

7.1. Trade-off Between Social Alignment and Driving Competence

The results across both environments reveal a consistent and monotonic trade-off between social alignment and driving competence. In the intersection environment, as the target angle increases from egoistic to altruistic, the collision rate decreases (29% → 23% → 11%), but so does the arrival rate (55% → 46% → 1%), while time stopped increases (21% → 33% → 71%) and speed decreases (20 → 16 → 3.8 km/h). The gridworld exhibits the same pattern: the egoistic agent crosses the intersection at full speed, the prosocial agent yields selectively, and the altruistic agent never crosses.

This trade-off is not merely a consequence of the regularizer making the agent more cautious. It reflects a genuine shift in the kind of behaviour the agent exhibits. The regularizer pushes the implicit reward of the egoistic agent to favour high r_{env} and low r_{social}, resulting in an agent that crosses as fast as possible, disrupting approaching traffic and accepting collisions as a cost of progress. On the other hand, the prosocial agent recovers an intermediate behavioural mode that balances individual and social objectives by stopping when cross-traffic is close but proceeding when safe. Finally, the altruistic agent prioritizes minimizing its social impact above all else, effectively withdrawing from the traffic interaction entirely to satisfy the objective of causing minimal disruption.

The prosocial regularization ($\alpha = 45°$) occupies a favourable position in this trade-off space, as it achieves a meaningful reduction in collision rate relative to the egoistic agent (23% vs 29%) while maintaining a reasonable arrival rate (46% vs 55%). This suggests that in the tested environments, moderate social alignment that is neither purely self-interested nor excessively yielding, produces the most practically useful policy when learning from heterogeneous demonstrations. The system designer can choose where on this spectrum to operate based on the deployment context, and the regularization strength λ provides the mechanism to do so.

This choice, however, is not purely technical. Selecting α_{target} determines whose interests the vehicle prioritizes in a shared environment. This decision has direct consequences for every surrounding road user. Our framework makes this decision explicit and measurable. The recovered $\hat{\alpha}$ provides a concrete quantity against which a deployed system's actual social behaviour can be verified. Under growing regulatory scrutiny, the auditability of an autonomous vehicle's social alignment is just as important as its controllability.

7.2. The Regularizer Does Not Replicate Experts

A noticeable finding across the intersection experiments is that the regularized IQ-Learn agents are systematically different from their corresponding experts. They are not simply imitations but distinct policies shaped by two different pressures.

The prosocial case is the most revealing. The IQ Pro-Reg agent achieves a lower collision rate than the prosocial expert it was trained to approximate (23% ±2.5% vs 25% ±1.5%), despite having access only to a mixed dataset where the prosocial demonstrations comprise one-third of the data. This improvement arises because the regularizer steers the social character of the reward, while the IQ-Learn objective absorbs physical driving dynamics from all demonstrations, including the cautious collision-avoidance patterns present in the altruistic data. The resulting policy inherits driving skill from the full dataset while expressing the social profile of a single target mode. This separation of physical competence from social intent is the core design principle of the framework, and the prosocial result provides direct empirical evidence that it works as intended.

The egoistic case illustrates the other side of this separation. The IQ Ego-Reg agent is more cautious than the egoistic expert: it drives at 20 km/h versus 28 km/h and achieves 55% arrival versus 70%. The gap reflects the fundamental difference between optimizing a known reward directly (what the expert does) and recovering a reward from mixed demonstrations (what IQ-Learn does). The IQ-Learn agent has seen cooperative and altruistic driving patterns alongside egoistic ones, and the regularization term in IQ-Learn itself penalizes extreme reward values. The result is an agent that is socially egoistic but physically more moderate than a pure egoistic expert.

These observations suggest that the framework does not simply filter the dataset for demonstrations matching the target angle. Instead, it constructs a new policy that combines the social profile specified by α_{target} with the physical driving competence available across the entire dataset. This is a feature of the reward regularization approach: because all transitions contribute to the IQ-Learn loss regardless of social alignment, the recovered Q-function encodes driving dynamics from all experts, and the MSE regularizer selects the social interpretation of those dynamics.

7.3. Baseline Collapse and the Role of Reward Design

In the 7×7 gridworld, the unregularized baseline and the altruistic-regularized agent converge to a non-driving policy, permanently stopping before the intersection. However, in the continuous intersection-v1 environment, the unregularized baseline does not collapse to a stop. Instead, it settles into a cautious profile that aligns closest with prosocial behaviour.

This reveals an important interaction between the IQ-Learn objective and the regularizer. Without an explicit social prior, the IQ-Learn objective attempts to explain the mixed demonstrations by randomly locking onto one of them. In the gridworld, this manifests as a cautious, aimless policy that settles on stopping, as it safely satisfies the frequent yielding seen in both altruistic and prosocial data without preferring the assertive, egoistic interpretation. In the more complex intersection-v1 environment, this averaging results in a hesitant but active policy that mirrors prosocial traits. Crucially, in both domains, the unregularized baseline lacks explicit directional intent. The regularizers are required to pull the agent out of these default local optima, whether that optimum is a complete stop or a hesitant prosocial compromise, and produce distinct targeted driving policies from the exact same mixed dataset. This demonstrates that the collapse is not caused by the environment reward function, but by the IQ-Learn objective lacking directional guidance when faced with ambiguous data, exactly the gap the regularizer is designed to fill.

Ultimately, this dynamic highlights a broader principle for IRL applications. While a regularizer provides critical directional guidance through ambiguous data, the base reward landscape dictates how aggressively that prior must fight to produce active task completion. If the underlying environment or objective inherently biases toward a specific default state a significant portion of the regularizer's

influence is consumed just to override it. A well-calibrated reward structure, by contrast, ensures that basic task completion is naturally viable. This allows the regularizer to focus entirely on its primary role - differentiating between the nuanced social modes of the expert demonstrations.

7.4. Robustness to Environment Complexity

The three experimental environments used in this paper represent a deliberate progression in structural complexity. The preliminary proof-of-concept environment provides a minimal, deterministic state space that isolates the core mechanics of the regularizer against exact ground-truth behaviours. The larger 7×7 gridworld introduces multi-agent stochasticity and discrete speed choices while remaining analytically tractable. Finally, the intersection environment scales the problem to a 70-dimensional continuous observation space with naturalistic traffic dynamics, requiring deep neural network approximation.

Despite these escalating complexities, the behavioural findings remain remarkably consistent across all three domains. First, the unregularized baseline reliably defaults to a highly cautious policy in every environment, whether that manifests as holding position until the other car passes (proof-of-concept), permanently yielding to all cross-traffic (7×7 gridworld), or adopting a hesitant, prosocial driving style (intersection). Second, the regularizer successfully manages to align the recovered reward with the target social profile across the board, scaling predictably with λ.

These consistent findings indicate that, across the tested environments, the framework's mechanisms to untangle ambiguous data and separate physical driving competence from social intent are not artefacts of a specific observation space or model size. Within the scope of these experiments, the framework's core properties consistently reproduce across settings of substantially varying complexity. These results open the door to further validation on more realistic simulators or naturalistic driving datasets, as discussed in Section 8.

8

Limitations and Future Works

This section discusses core limitations of the current work and suggests directions for future research.

8.1. Synthetic Expert Populations

The biggest limitation of this work is that all experiments rely on synthetic experts trained in a simulator, rather than naturalistic driving data. This choice was necessary due to real-world datasets lacking ground truth social intent. On the other hand, synthetic experts provide a controlled setting where the underlying objective is explicitly defined and the recovered $\hat{\alpha}$ can be compared against a known ground truth.

However, this means that the framework has only been validated in environments where r_{env} and r_{social} are explicitly computable from the simulator state. Closing the sim-to-real gap remains a fundamental challenge. Possible approaches include unsupervised discovery of social modes from naturalistic trajectories to initialize the prior, or transfer learning where the regularizer is calibrated in simulation and fine tuned on real data.

8.2. Hand-Specified Reward Decomposition

Although motivated by the difficulty of hand-crafting reward functions, this framework still requires the designer to manually specify r_{env} and r_{social}. While the IQ-Learn objective implicitly recovers the core driving reward from demonstrations, the social decomposition that steers it remains a hand-crafted specification. This creates an inherent tension because replacing a complete reward function with a semantically meaningful social signal simply shifts the engineering burden, demanding specific domain knowledge regarding influence zones, progress measures, and disruption criteria.

A particularly clear instance of this hand-crafting burden is the influence zone itself, which is currently specified per environment based on the designer's knowledge of where ego actions can affect neighbour behaviour. Defining this region requires non-trivial familiarity with the environment's geometry, traffic priority structure, and the spatial reach of vehicle interactions, none of which transfer automatically across road topologies. Learning the influence zone directly from observed counterfactual interactions would remove this manual specification and complement the broader effort to derive r_{social} from data rather than design.

Despite these constraints, the crucial distinction is that r_{env} and r_{social} encode structural social intent rather than exhaustive driving objectives. Our results confirm that this partial specification is sufficient to resolve behavioural ambiguity without constraining the physical driving competence recovered from the data. Nevertheless, eliminating this manual step entirely remains an open problem. Future research could explore learning the social decomposition directly from multi-agent trajectory data

using observational proxy measures like time-to-collision or yielding indicators, or by grounding the decomposition in causal responsibility metrics that bypass the need for environment-specific engineering.

8.3. Absence of Malicious Demonstrations

The introduction motivates the framework partly through the need to handle datasets containing malicious behaviour. However, the experiments only test mode alignment among not explicitly adversarial modes. The originally planned contaminated dataset experiments were not conducted due to time constraints.

This, however, is still important to explore. While the egoistic agent does produce aggressive behaviour, this is qualitatively different from the active disruption that a purposefully malicious agent would cause. The ego-centric formulation naturally accommodates adversarial behaviour and extending the evaluation to malicious contamination is a natural and very important next step.

8.4. Discrete Action Space and Environment Scope

All experimental environments use discrete actions. Real autonomous driving, however, involves continuous control over acceleration and steering. IQ-Learn's original implementation naturally supports this transition with an actor-critic architecture. While this extension was not explored in the current work, the reward regularization mechanism itself is fundamentally action-space agnostic. Therefore, the theoretical framework should transfer directly to continuous domains. The practical stability and performance in those settings still remain unvalidated.

8.5. Fixed Target Angle

The current framework requires the designer to specify α_{target} before training. There is no mechanism to discover the appropriate social profile from data or to adapt the target during training. Although the framework reliably enforces a specified social profile, the choice of α_{target} dictates the resulting behavioural mode and the designer may not know the appropriate angle for a given deployment task.

Future work could address this through adaptive target selection. One idea could be using a validation set of desired behaviours to search over angles, or jointly optimizing α alongside the Q-function. A more aspiring direction could extend the framework to recover multiple social modes simultaneously, combining the regularizer with mixture-of-experts models rather than selecting a single mode.

8.6. Scope of the Social Impact Signal

The current formulation of r_{social} aggregates the ego vehicle's influence on neighbouring vehicular agents at the conflict point. In real world traffic, however, the set of agents affected by the ego vehicle's actions extends well beyond other vehicles: pedestrians, cyclists, and wheelchair users routinely share intersections and shared zones with cars, and the framework as currently defined does not represent them in the regularizer.

Extending r_{social} to incorporate non-vehicular road users raises two technical questions. First, the social impact signal must be redefined to handle various neighbour types whose state representation, kinematic constraints, and disruption metrics differ substantially from those of cars: yielding to a pedestrian at a crosswalk, passing a cyclist, and approaching a vehicle at an intersection each represent distinct forms of impact. Second, the relative weighting of these neighbour classes becomes a design choice of its own: a uniform aggregation would conflate substantial differences in vulnerability, while per class weights would introduce additional hand specified parameters into a framework that already requires the designer to specify r_{env} and r_{social}.

A practical first step would be to define class-conditioned impact terms, for example

$$r_{\text{social}}^{\text{vehicular}}$$

and

$$r_{\text{social}}^{\text{vulnerable}}$$

, and to combine them through explicit, inspectable weights. While it introduces an additional set of designer specified parameters, it ensures that the relative priority assigned to vulnerable road users remains an auditable quantity. A more ambitious direction could replace these manual weights with values derived directly from observed yielding behaviour in naturalistic datasets, or ground them in broader causal responsibility metrics. This would eliminate the need for manual specification.

9

Social Implications

The preceding chapters established the technical contributions of the Social Impact Regularized IQ-Learn. The implications of this framework, however, extend beyond its algorithmic properties. Exposing the social orientation of a learned policy as an inspectable parameter raises new questions about how that parameter should be set, who has the legitimacy to set it, and what safeguards should govern its use.

This chapter situates the framework within its broader societal context. Section 9.1 positions the work within the value alignment literature, framing the shift from implicit learning to explicit social design. Section 9.2 connects this shift to broader efforts in algorithmic accountability and emerging AI regulation. The remaining sections look at the questions that this transparency raises: how the framework affects different stakeholder groups (Section 9.3), how a static social prior interacts with evolving traffic ecosystems (Section 9.4), how the same tunability that enables accountability also enables deliberate misuse (Section 9.5), and the unresolved normative question of whose values should govern the chosen social orientation (Section 9.6).

9.1. From Implicit Learning to Explicit Social Design

A persistent gap in the value alignment literature is the translation of abstract alignment principles into concrete, inspectable system parameters [39, 40]. The results presented in this thesis represent one step in this direction. By demonstrating that the same mixed dataset yields distinct egoistic, prosocial and altruistic policies depending on a single inspectable parameter, the proposed framework moves social alignment from a side effect of training data into a deliberate architectural choice.

The field of machine learning for autonomous systems has historically prioritized performance optimization, allowing agents to navigate complex environments efficiently [1, 7]. As the application of these systems has advanced, a consistent limitation has become apparent that systems optimized across heterogeneous data risk inadvertently encoding conflicting implicit goals, producing behaviours that are difficult to anticipate and even harder to audit [6, 7, 39, 41]. The approach presented in this thesis addresses this challenge not by discarding data driven learning, but by separating what is learned from data (physical driving competence) from what is chosen by design (social intent). In doing so, it contributes directly to the pursuit of interpretable artificial intelligence.

This connects directly to an ongoing challenge in artificial intelligence research: value alignment [40]. The core problem is characterized as systems that pursue objectives which are imperfectly specified with respect to what their designers actually intend, a problem that has become increasingly significant recently [39]. Furthermore, because the normative and technical aspects of AI alignment are deeply interconnected, a crucial distinction emerges between systems that simply imitate human behaviour and those explicitly designed to uphold chosen values [40]. The framework proposed here puts this distinction into practice. While IQ-Learn extracts baseline driving skills from the dataset, the regularizer

actively steers the system's social behaviour. This setup ensures that human values are explicitly built into the agent's actions, rather than leaving its social dynamics to chance [6].

9.2. Algorithmic Accountability and the Broader Alignment Landscape

The architectural shift demonstrated above aligns closely with broader efforts in algorithmic accountability. Reinforcement Learning from Human Feedback (RLHF) [42] and Constitutional AI [43] both face an analogous problem: a base model trained on a large, mixed corpus must be steered toward a chosen interactive profile that the data alone does not specify. RLHF accomplishes this through learned preference models, whereas Constitutional AI employs an explicit set of written principles. The framework proposed here occupies a related position within inverse reinforcement learning with the social function playing a role analogous to a constitution or a learned preference model. Although the mechanism differs, the core design philosophy is shared: general capability is derived from data, while behavioural disposition is derived from explicit specification.

A similar paradigm is emerging within contemporary regulation. Beyond the EU AI Act's classification of autonomous vehicle systems as high-risk AI and the ongoing debates over the governance of their social behaviour [44, 45], the wider regulatory environment is converging on requirements for auditable, documented AI. NIST's AI Risk Management Framework and ISO/IEC 42001's certification standard for AI management systems [46, 47] both emphasise two complementary properties: that a deployed AI system's behavioural commitments be inspectable up front, and that its actual behaviour be verifiable against those commitments. A learned reward function lacking an explicit social dimension is structurally ill-suited to meet these requirements, since there are no documented behavioural commitments for an audit to check against. The regularized approach instead maps cleanly onto these requirements with α_{target} documenting the chosen profile up front (inspectable), and the recovered $\hat{\alpha}$ providing the post-hoc verification that the deployed policy embodies it (verifiable).

Ultimately, meeting these regulatory demands transforms how we handle algorithmic accountability. When a conventional, opaque model makes a detrimental interactive choice on the road, fault is notoriously difficult to trace [44]. By formalizing interactive tendencies into a transparent metric, the conversation shifts from reactive fault attribution after an incident to proactive policy making, grounding discussions of public safety in auditable parameters rather than behavioural speculation. Transparency, however, remains a necessary but not sufficient condition for accountability: making α explicit reveals exactly what social profile was chosen, but does not in itself ensure that the chosen profile is appropriate for all road contexts, users, or stakeholders.

9.3. Stakeholder Perspectives

Addressing this gap requires moving beyond technical transparency and acknowledging the diversity of the traffic environment. When we assign a single value to α_{target}, we inadvertently treat the public as a singular stakeholder. It is not. The interests of distinct groups affected by this choice are fundamentally different, presenting trade-offs that cannot be solved through algorithmic optimization.

The most consequential of these choices concerns vulnerable road users. The social impact signal in this work captures how the ego vehicle's actions affect other vehicular agents at the conflict point. Pedestrians, cyclists, wheelchair users, and other non-vehicular road users are not represented in the regularizer. As a result, the current framework possesses a substantial blind spot regarding the safety and prioritization of the most vulnerable road users. This blind spot creates severe behavioural trade-offs: an altruistic agent that withdraws from a traffic interaction entirely is helpful to surrounding cars but may be obstructive to a pedestrian waiting to cross, while an egoistic agent that asserts its right of way against approaching vehicles may assert it more readily against a cyclist or a wheelchair user. Deciding whose safety counts in r_{social}, and how to balance these competing interests, is a deeply normative decision that the current scalar α fails to capture.

Even within vehicular interactions, relevant stakeholders rarely share a unified objective. A passenger utilizing a mobility service may favour an egoistic profile that minimizes trip time, whereas a regulator may permit only a prosocial baseline. Automated fleet operators find themselves simultaneously accountable to passengers, regulators, and the broader public - three distinct groups whose expectations regarding the appropriate α do not naturally align.

This multi-stakeholder tension is directly reflected in our empirical results within the intersection-v1 environment. Here, the choice of α directly dictated the collision rates experienced by surrounding traffic, yielding 29%, 23%, and 11% under the egoistic, prosocial, and altruistic configurations, respectively, alongside corresponding arrival rates of 55%, 46%, and 1%. These performance trade-offs reveal the conflicting incentives: a passenger prioritizing transit efficiency has a clear interest in the egoistic profile, while a regulator concerned with public safety has equally clear grounds to mandate the prosocial baseline, sacrificing nine percentage points of throughput to achieve a six-point reduction in collisions. An operator accountable to both parties must justify its operational choices somewhere on this exposed trade-off curve.

Crucially, while the proposed framework provides the single dial required to navigate these trade-offs, it remains neutral on who should control it. What it does change is that the choice becomes attributable. Different α values imply systematically different collision rates, but in conventional models this variation is buried in the policy and cannot cleanly be traced to a design choice. Because α is documented at deployment and the recovered $\hat{\alpha}$ can be audited post-hoc, the social profile becomes an explicit decision. Two consequences follow. First, liability exposure can be tied to that declared choice: operating under an egoistic α carries a known, attributable collision rate increase, making the distribution of risk among operator, manufacturer, and surrounding public explicit and contestable. Second, social alignment becomes an insurable metric supporting risk-adjusted insurance premiums, introducing a quantifiable market mechanism entirely absent from current autonomous vehicle liability discussions.

9.4. Population Level Effects and the Limits of a Static Prior

Considering the framework's application at scale introduces further implications. It has been demonstrated that driving constitutes a sequence of social dilemmas in which individually rational behaviour (acting egoistically) produces collectively suboptimal outcomes such as increased congestion and reduced cooperative throughput [19]. Viewed through this game-theoretic lens, the ability to certify a fleet's social orientation offers a concrete mechanism to mitigate these dilemmas at scale. A municipal regulator could, in principle, mandate a minimum prosociality threshold for autonomous vehicles operating within dense urban centers, effectively suppressing the free-rider dynamics that emerge when individual agents optimize purely for personal progress [19, 45].

At the same time, widespread deployment of uniformly prosocial vehicles introduces its own risks. If a large fraction of road users are certified as prosocial while a minority of human-driven or differently configured vehicles remain egoistic, the prosocial agents may be systematically exploited, yielding consistently while egoistic agents advance, mirroring the tragedy of the commons dynamics identified in heterogeneous traffic [48]. The framework therefore raises a regulatory question that extends beyond individual vehicle certification: who sets the population-level distribution of social profiles, and how is that distribution enforced and monitored across a heterogeneous fleet?

A second limitation arises regarding the static nature of the social target. The current methodology relies on a fixed social target prior to deployment. However, empirical evidence consistently shows that driving norms are highly context dependent and vary significantly across geographic regions and road environments [49, 50]. What is considered an appropriate distance gap in a shared urban zone may be perceived as dangerously aggressive on a motorway, and vice versa. Across cultures, acceptable driving behaviours differ substantially, with documented variation in violation rates, yielding behaviour, and attitudes toward assertive manoeuvres. [49]. Large scale cross cultural surveys of autonomous vehicle

decision making preferences further confirm that no single behavioural profile commands universal acceptance [50]. A static social parameter cannot account for this contextual and cultural variation, raising the question of how a certification framework might remain meaningful while accommodating the behavioural flexibility that real-world traffic demands.

9.5. Risks of an Unregulated Parameter

Making behavioural alignment an explicit, adjustable parameter exposes it to deliberate misalignment. The same mechanism that enables the specification of prosocial, cooperative behaviour could equally be used to instantiate aggressive, egoistic policies if a designer prioritizes individual transit speed above collective safety, and the framework provides no internal mechanism to resist this.

The structural incentives to engage in such manipulation are clear. A logistics operator under competitive pressure on delivery times has a direct economic benefit to push α toward the egoistic end, particularly in jurisdictions without a mandated prosociality minimum. The resulting increase in collision risk is not borne by the operator but externalized onto the surrounding traffic. A manufacturer could pursue the same configuration under the framing of feature differentiation, such as a sport or performance driving profile, converting the social parameter into a product option and transferring the safety trade-off to the consumer. These are not edge cases but foreseeable commercial responses to an unregulated parameter.

The risk also operates across regulatory boundaries. If one jurisdiction certifies prosocial defaults while a neighbouring one does not, vehicles registered externally may operate egoistically within the regulated zone, systematically exploiting the compliant agents around them. This is less a failure of the framework itself than an illustration of how its effectiveness is dependent on coordinated adoption, raising the question of what enforcement architecture would be required to make certification robust against jurisdictional exploitation.

9.6. Whose Choice is the Target Social Alignment?

The limitations identified throughout this chapter all come down to a single underlying problem: the framework makes social alignment technically tractable but leaves the normative question of whose values should govern that alignment entirely unresolved. Determining the appropriate prosociality threshold is not a calibration decision but a political one, involving trade-offs between individual operators, commercial interests, and the broader travelling public whose safety is directly affected by choices they have no part in making [40].

This is not a limitation unique to the framework, but it is sharpened by it in a productive way. By making α an adjustable parameter, the framework converts a question that was previously buried in training data into one that must be answered openly. That visibility is necessary for meaningful oversight: regulators cannot govern what they cannot see, and the public cannot contest boundaries that are never made legible. The framework does not resolve the question of whose values should prevail, but it ensures that the question can at least be asked.

10

Conclusion

This work addresses a known problem in inverse reinforcement learning for autonomous driving: when demonstrations come from drivers with conflicting social preferences, standard IRL has no directional intent that helps it to choose among the vast number of reward functions that fit the data, and the recovered policy collapses onto an arbitrary mode. The Social Impact Regularized IQ-Learn framework introduces a Tikhonov-style regularizer, parameterized by a social signal. The IQ-Learn objective recovers physical driving competence from all demonstrations, while the regularizer selects the social interpretation. Empirical evaluations progressing from deterministic gridworlds to a continuous observation intersection simulator confirmed the effectiveness of this approach. The results demonstrated that a single mixed dataset can reliably yield distinctly separated behavioural profiles, ranging from egoistic to altruistic, based entirely on the specified regularization target. Crucially, the empirical analysis quantified a persistent, measurable trade-off between physical driving efficiency and social yielding. This confirms that the regularizer successfully reconstructs a targeted policy that balances individual progress with social impact, rather than merely imitating a subset of the expert data.

Beyond autonomous driving, this methodology contributes directly to the broader challenge of AI value alignment by demonstrating how abstract alignment principles can be translated into concrete, inspectable system parameters. By shifting the paradigm from implicit behavioural learning to deliberate social specification, this approach offers a structural blueprint for creating transparent and auditable systems in interactive environments.

While this methodology provides a robust foundation, bridging the gap to real-world deployment requires further technical exploration. Future research must evaluate the framework on naturalistic human driving datasets to close the simulation gap. Furthermore, expanding the scope of the social impact signal to account for pedestrians, cyclists, and other non-vehicular road users is necessary to ensure the system's safety metrics reflect the full complexity of driving environments.

Yet, even if these technical challenges are resolved, a more fundamental question of values remains. This framework enforces a chosen social profile but does not address who should choose it or whether one profile is appropriate across different road contexts and groups of road users. Rather than an algorithmic limitation, this opens the door for a vital public dialogue, inviting regulators, manufacturers, and society at large to collaboratively chart the safest course ahead.

References

[1] Wenshuo Wang et al. "Social Interactions for Autonomous Driving: A Review and Perspectives". In: *Foundations and Trends® in Robotics* 10.3–4 (Nov. 2022), pp. 198–377. ISSN: 1935-8261. DOI: 10.1561/2300000078. URL: http://dx.doi.org/10.1561/2300000078.

[2] Dorsa Sadigh et al. "Planning for Autonomous Cars that Leverage Effects on Human Actions". In: *Robotics: Science and Systems*. 2016. URL: https://api.semanticscholar.org/CorpusID:7087988.

[3] G. Markkula et al. "Defining interactions: a conceptual framework for understanding interactive behaviour in human and automated road traffic". In: *Theoretical Issues in Ergonomics Science* 21.6 (2020), pp. 728–752. DOI: 10.1080/1463922X.2020.1736686. eprint: https://doi.org/10.1080/1463922X.2020.1736686. URL: https://doi.org/10.1080/1463922X.2020.1736686.

[4] Xu Chen, Xuan Di, and Zechu Li. "Social Learning for Sequential Driving Dilemmas". In: *Games* 14.3 (May 2023), p. 41. DOI: 10.3390/g14030041.

[5] Benjamin Eysenbach, Sergey Levine, and Ruslan Salakhutdinov. *Replacing Rewards with Examples: Example-Based Policy Search via Recursive Classification*. 2021. arXiv: 2103.12656 [cs.LG]. URL: https://arxiv.org/abs/2103.12656.

[6] Dario Amodei et al. *Concrete Problems in AI Safety*. 2016. arXiv: 1606.06565 [cs.AI]. URL: https://arxiv.org/abs/1606.06565.

[7] W. Bradley Knox et al. "Reward (Mis)design for autonomous driving". In: *Artificial Intelligence* 316 (2023), p. 103829. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2022.103829. URL: https://www.sciencedirect.com/science/article/pii/S0004370222001692.

[8] Andrew Ng and Stuart Russell. "Algorithms for Inverse Reinforcement Learning". In: *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning* (May 2000).

[9] Pieter Abbeel and Andrew Y. Ng. "Apprenticeship learning via inverse reinforcement learning". In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: Association for Computing Machinery, 2004, p. 1. ISBN: 1581138385. DOI: 10.1145/1015330.1015430. URL: https://doi-org.tudelft.idm.oclc.org/10.1145/1015330.1015430.

[10] Zhiyu Huang, Jingda Wu, and Chen Lv. *Driving Behavior Modeling using Naturalistic Human Driving Data with Inverse Reinforcement Learning*. 2021. arXiv: 2010.03118 [cs.RO]. URL: https://arxiv.org/abs/2010.03118.

[11] Sascha Rosbach et al. "Driving with style: Inverse reinforcement learning in general-purpose planning for automated driving". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 2658–2665.

[12] Monica Babeş-Vroman et al. "Apprenticeship learning about multiple intentions". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 897–904. ISBN: 9781450306195.

[13] Jaedeug Choi and Kee-eung Kim. "Nonparametric Bayesian Inverse Reinforcement Learning for Multiple Reward Functions". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/140f6969d5213fd0ece03148e62e461e-Paper.pdf.

[14] Jeremy Morton and Mykel J. Kochenderfer. *Simultaneous Policy Learning and Latent State Inference for Imitating Driver Behavior*. 2017. arXiv: 1704.05566 [cs.LG]. URL: https://arxiv.org/abs/1704.05566.

[15] Saurabh Arora and Prashant Doshi. "A survey of inverse reinforcement learning: Challenges, methods and progress". In: *Artificial Intelligence* 297 (2021), p. 103500. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2021.103500. URL: https://www.sciencedirect.com/science/article/pii/S0004370221000515.

[16] Wim B. G. Liebrand and Charles G. McClintock. "The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation". In: *European Journal of Personality* 2.3 (1988), pp. 217–230. DOI: 10.1002/per.2410020304.

[17] Liting Sun et al. *Courteous Autonomous Cars*. 2018. arXiv: 1808.02633 [cs.RO]. URL: https://arxiv.org/abs/1808.02633.

[18] Jack Geary and Henry Gouk. *Altruistic Decision-Making for Autonomous Driving with Sparse Rewards*. 2020. arXiv: 2007.07182 [cs.GT]. URL: https://arxiv.org/abs/2007.07182.

[19] Wilko Schwarting et al. "Social behavior for autonomous vehicles". In: *Proceedings of the National Academy of Sciences* 116.50 (2019), pp. 24972–24978.

[20] Michal Monselise and Christopher C. Yang. "Detecting aggressive driving patterns in drivers using vehicle sensor data". In: *Transportation Research Interdisciplinary Perspectives* 14 (2022), p. 100625. ISSN: 2590-1982. DOI: https://doi.org/10.1016/j.trip.2022.100625. URL: https://www.sciencedirect.com/science/article/pii/S2590198222000872.

[21] Jordanka Kovaceva, Irene Isaksson-Hellman, and Nikolce Murgovski. "Identification of aggressive driving from naturalistic data in car-following situations". In: *Journal of Safety Research* 73 (2020), pp. 225–234. ISSN: 0022-4375. DOI: https://doi.org/10.1016/j.jsr.2020.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0022437520300335.

[22] Divyansh Garg et al. *IQ-Learn: Inverse soft-Q Learning for Imitation*. 2022. arXiv: 2106.12142 [cs.LG]. URL: https://arxiv.org/abs/2106.12142.

[23] Brian D Ziebart et al. "Maximum entropy inverse reinforcement learning". In: *AAAI Conference on Artificial Intelligence*. 2008, pp. 1433–1438.

[24] A. N. Tikhonov. "Solution of Incorrectly Formulated Problems and the Regularization Method". In: *Soviet Mathematics Doklady* 4 (1963). English translation of Dokl. Akad. Nauk SSSR, 151, 501–504, pp. 1035–1038.

[25] Federal Highway Administration. *Next Generation Simulation (NGSIM) Vehicle Trajectory and Supporting Data*. Tech. rep. FHWA-HRT-07-030. U.S. Department of Transportation, 2007. URL: https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm.

[26] Robert Krajewski et al. *The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems*. 2018. arXiv: 1810.05642 [cs.CV]. URL: https://arxiv.org/abs/1810.05642.

[27] Pei Sun et al. *Scalability in Perception for Autonomous Driving: Waymo Open Dataset*. 2020. arXiv: 1912.04838 [cs.CV]. URL: https://arxiv.org/abs/1912.04838.

[28] Songyuan Zhang et al. *Confidence-Aware Imitation Learning from Demonstrations with Varying Optimality*. 2022. arXiv: 2110.14754 [cs.LG]. URL: https://arxiv.org/abs/2110.14754.

[29] Daniel S. Brown et al. *Extrapolating Beyond Suboptimal Demonstrations via Inverse Reinforcement Learning from Observations*. 2019. arXiv: 1904.06387 [cs.LG]. URL: https://arxiv.org/abs/1904.06387.

[30] Daniel S. Brown, Wonjoon Goo, and Scott Niekum. *Better-than-Demonstrator Imitation Learning via Automatically-Ranked Demonstrations*. 2019. arXiv: 1907.03976 [cs.LG]. URL: https://arxiv.org/abs/1907.03976.

[31] Antonio Mone, Frans A. Oliehoek, and Luciano Cavalcante Siebert. *CoMI-IRL: Contrastive Multi-Intention Inverse Reinforcement Learning*. 2026. arXiv: 2602.07496 [cs.LG]. URL: https://arxiv.org/abs/2602.07496.

[32] Ran Tian, Masayoshi Tomizuka, and Liting Sun. *Learning Human Rewards by Inferring Their Latent Intelligence Levels in Multi-Agent Games: A Theory-of-Mind Approach with Application to Driving Data*. 2021. arXiv: 2103.04289 [cs.AI]. URL: https://arxiv.org/abs/2103.04289.

[33] Ashwin George et al. "Feasible Action-Space Reduction as a Metric of Causal Responsibility in Multi-Agent Spatial Interactions". In: *ECAI 2023*. IOS Press, Sept. 2023. ISBN: 9781643684376. DOI: 10.3233/faia230349. URL: http://dx.doi.org/10.3233/FAIA230349.

[34] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. *On a Formal Model of Safe and Scalable Self-driving Cars*. 2018. arXiv: 1708.06374 [cs.RO]. URL: https://arxiv.org/abs/1708.06374.

[35] Rodolfo Valiente et al. *Learning-based social coordination to improve safety and robustness of cooperative autonomous vehicles in mixed traffic*. 2022. arXiv: 2211.11963 [cs.RO]. URL: https://arxiv.org/abs/2211.11963.

[36] Deepak Ramachandran and Eyal Amir. "Bayesian inverse reinforcement learning". In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI'07. Hyderabad, India: Morgan Kaufmann Publishers Inc., 2007, pp. 2586–2591.

[37] Edouard Leurent. *An Environment for Autonomous Driving Decision-Making*. https://github.com/eleurent/highway-env. 2018.

[38] Arne Kesting, Martin Treiber, and Dirk Helbing. *Agents for Traffic Simulation*. 2008. arXiv: 0805.0300 [physics.soc-ph]. URL: https://arxiv.org/abs/0805.0300.

[39] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group, 2019. ISBN: 9780525558620. URL: https://books.google.nl/books?id=M1eFDwAAQBAJ.

[40] Iason Gabriel. "Artificial Intelligence, Values, and Alignment". In: *Minds and Machines* 30.3 (2020), pp. 411–437. ISSN: 1572-8641. DOI: 10.1007/s11023-020-09539-2. URL: http://dx.doi.org/10.1007/s11023-020-09539-2.

[41] Joar Skalse and Alessandro Abate. *Partial Identifiability and Misspecification in Inverse Reinforcement Learning*. 2024. arXiv: 2411.15951 [cs.LG]. URL: https://arxiv.org/abs/2411.15951.

[42] Paul Christiano et al. "Deep reinforcement learning from human preferences". In: (June 2017). DOI: 10.48550/arXiv.1706.03741.

[43] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback*. 2022. arXiv: 2212.08073 [cs.CL]. URL: https://arxiv.org/abs/2212.08073.

[44] European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Entered into force 1 August 2024. July 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[45] Patricia Derler, Noushin Mehdipour, and Radboud Tebbens. "How should autonomous vehicles drive? Policy, methodological, and social considerations for designing a driver". In: *Humanities and Social Sciences Communications* 9 (Aug. 2022). DOI: 10.1057/s41599-022-01286-2.

[46] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC 42001:2023, Information Technology — Artificial Intelligence — Management System*. International Standard. Geneva, Switzerland: ISO/IEC, Dec. 2023. URL: https://www.iso.org/standard/42001.

[47] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. NIST AI 100-1. U.S. Department of Commerce, 2023. DOI: 10.6028/NIST.AI.100-1. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[48] Ricardo Simão and Lucas Wardil. "Social dilemma in traffic with heterogeneous drivers". In: *Physica A: Statistical Mechanics and its Applications* 561 (2021), p. 125235. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2020.125235. URL: https://www.sciencedirect.com/science/article/pii/S0378437120306518.

[49] Türker Özkan et al. "Cross-cultural differences in driving behaviours: A comparison of six countries". In: *Transportation Research Part F-traffic Psychology and Behaviour - TRANSP RES PT F-TRAFFIC PSYCH* 9 (May 2006), pp. 227–242. DOI: 10.1016/j.trf.2006.01.002.

[50] Amilcar Gröschel Jr, Edmond Awad, and Jonathan Schulz. "2018 Article The Moral Machine". In: *Nature* (Jan. 2018).

[51] J. Cohen. *Statistical power analysis for the behavioral sciences (2nd ed.)* Lawrence Erlbaum, 1988.

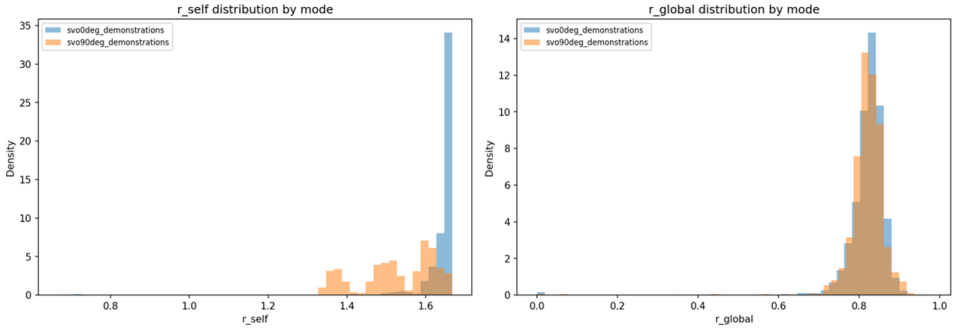
[52] Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. "Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping". In: *Proceedings of the Sixteenth International Conference on Machine Learning*. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 278–287. ISBN: 1558606122.

A

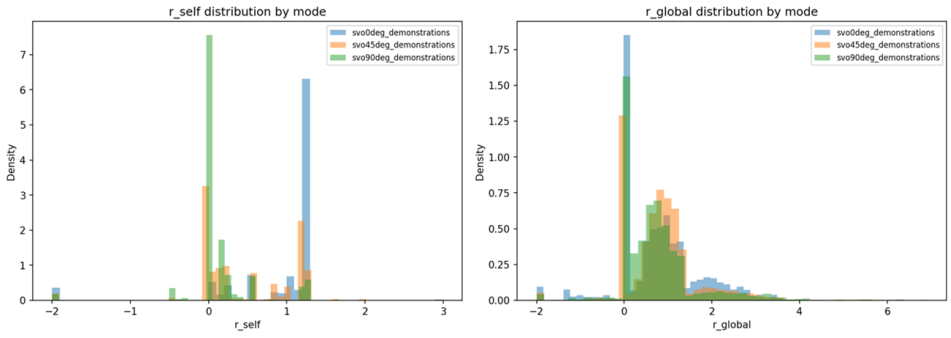
SVO Reward Analysis

A.1. Distinction from Classical SVO

The classical SVO formulation in the context of autonomous driving [19] defines r_{global} as the average reward of all traffic participants, aggregating what other agents experience as a measure of collective welfare. By making a preliminary validation, however, we find that this formulation leads to a collapse in separability (Figure A.1) and the resulting social scores fail to meaningfully distinguish between behavioural modes.



((a)) Comparison of two behavioural modes.



((b)) Comparison expanded to three behavioural modes.

Figure A.1: Distributions of self-reward (r_{self}) and global reward (r_{global}) components across different environment configurations.

Due to this, our formulation differs significantly. We redefine r_{global} not as the reward experienced by other vehicles, but as a measure of how the ego vehicle's actions affect others and call it r_{social}. This is an ego-centric social impact signal: it captures the consequences of the ego's behaviour on its neighbours,

rather than the neighbours' independent utilities. In the classical formulation, a passive ego vehicle that simply stays out of the way would receive a high r_{global} because other vehicles would achieve rewards independently. In our formulation, such a vehicle would receive a near-zero r_{social} because it exerts no influence, positive or negative, on its neighbours. Our r_{social} is only nonzero when the ego vehicle is in a position to meaningfully affect the surrounding traffic, which we implement through the concept of an influence zone.

Because our underlying r_{social} measures ego impact rather than group welfare, we refer to our framework as SVO inspired regularization. The angle α controls how much weight the regularizer places on the ego's social impact, not on the group's aggregate utility.

A.2. Reward Component Separability

The effectiveness of the social regularization depends on the assumption that r_{env} and r_{social} carry complementary information and that they vary independently enough across behavioural modes to allow R_{total} to discriminate between them. When this assumption is violated, the regularizer loses its discriminative power.

We empirically demonstrate this failure mode using the classical SVO formulation. To quantify this difference we analyse the reward component distributions across the three behavioural modes in the intersection-v1 dataset under both formulations. Table A.1 reports the per-episode cumulative returns.

Table A.1: Per-episode cumulative returns by behavioural mode and r_{global} formulation (intersection-v1). Values are reported as mean ± standard deviation.

	Classical		Ego-centric	
Mode	G_{self}	G_{global}	G_{env}	G_{social}
Egoistic	+8.17 ± 3.45	+7.00 ± 3.93	+7.13 ± 3.26	−2.00 ± 4.84
Prosocial	+6.76 ± 3.58	+12.16 ± 4.81	+4.88 ± 3.44	+0.68 ± 6.75
Altruistic	+2.47 ± 1.24	+12.24 ± 4.99	+0.01 ± 0.96	+1.76 ± 5.51

As observed in Table A.1, while the classical formulation yields distinct G_{self} returns across all modes, the global return plateaus between the Prosocial and Altruistic experts (+12.16 vs. +12.24). This lack of separation in cumulative returns stems directly from the heavily overlapping per-step rewards shown earlier in Figure A.1. Because the classical r_{global} fails to exhibit independent variance, the target rotation α effectively reweights signals that cannot be distinguished from one another. This structural limitation firmly motivates the ego-centric social impact formulation adopted in our primary methodology.

Table A.2 formalizes the pairwise separability using Cohen's d [51], which measures the effect size between two distributions as the difference in means divided by the pooled standard deviation. Values of $|d| < 0.2$ indicate negligible separation, 0.2–0.5 small, 0.5–0.8 medium, and > 0.8 large. Because the regularizer operates on cumulative returns (Section 5.3.1), both analyses are conducted at the episode level.

Table A.2: Pairwise Cohen's d on per-episode cumulative returns across behavioural modes (intersection-v1). A positive d indicates that the first-listed mode has the higher mean. The G_{global} columns compare the classical formulation (average neighbour reward) against the ego-centric formulation (social impact of the ego vehicle) adopted in this work.

		$d(G_{\text{global}})$	
Comparison	$d(G_{\text{self}})$	Classical	Ego-centric
Egoistic vs. Prosocial	+0.67	−1.17	−0.46
Prosocial vs. Altruistic	+1.93	−0.02	−0.17
Egoistic vs. Altruistic	+2.96	−1.16	−0.72

The G_{self} component shows medium-to-large separation across all pairs, confirming that the three expert policies produce measurably different driving behaviour at the episode level. The critical difference lies in G_{global}. Under the classical formulation, the previously noted plateau between prosocial and altruistic experts yields a negligible pairwise effect size of $d = -0.02$. The egoistic mode is well-separated from both ($|d| > 1.0$), but this separation is driven by the same mechanism as G_{self}: faster, more assertive driving produces different neighbour experiences regardless of social intent. The classical G_{global} can distinguish aggressive from non-aggressive driving, but cannot differentiate between the cooperative modes, precisely the distinction the regularizer must make when steering from prosocial to altruistic.

Under the ego-centric formulation, the three modes are monotonically ordered: G_{global} increases from -2.00 (egoistic) to $+0.68$ (prosocial) to $+1.76$ (altruistic). The prosocial–altruistic separation remains modest ($d = -0.17$), but all three pairwise effect sizes in Table A.2 are consistently negative, meaning that in every comparison the more socially oriented mode produces the higher G_{global}. This consistent sign is what the MSE regularizer exploits: because the ordering never flips, the per-episode gradient of the MSE loss always points toward higher G_{global} for more altruistic targets. Even though the gradient is small in magnitude, it accumulates over training in a single direction, allowing the regularizer to steer the recovered reward toward the target social profile. By contrast, the collapsed classical signal provides no such informative gradient between the cooperative modes, explaining its loss of discriminative power.

B

Alternative Integration Strategies

This section details two alternative methods for integrating the social prior into the IQ-Learn framework that were considered but ultimately set aside in favour of the Reward Regularization approach (Section 4.3.1).

B.1. Bellman Target Shift

The first integration strategy modifies the implicit reward seen by Term 1 of the IQ-Learn objective. Instead of evaluating $\hat{r}$ directly, the Bellman target is shifted by the R_{total} signal:

$$\hat{r}_{\text{shifted}}(s, a) = \hat{r}(s, a) - \lambda \cdot R_{\text{total}}(s, a \mid \alpha_{\text{target}}). \tag{B.1}$$

This shifted reward replaces $\hat{r}$ in Term 1, so the objective now rewards expert transitions whose implicit reward exceeds the R_{total} target. The intuition is that the Q-function should assign higher reward to transitions that are both physically plausible (high $\hat{r}$) and socially aligned (high R_{total}). While this can be interpreted as potential-based reward shaping [52], which preserves optimal-policy invariance in the forward RL setting, the guarantee does not straightforwardly transfer to IRL. Here we are recovering the reward, not optimizing a known one, so shifting the Bellman target changes what reward the Q-function must explain rather than merely reshaping the learning signal.

This approach also introduces two mathematical complications. Term 3 of the IQ-Learn loss computes $\frac{1}{4\alpha_\chi}\mathbb{E}\left[\hat{r}^2\right]$, which bounds the magnitude of the learned implicit reward as established by [22]. If the shifted reward $\hat{r}_{shifted}$ were used in Term 3, the regularizer would penalize large values of $\lambda \cdot R_{\text{total}}$, which is a fixed external quantity, rather than large Q-values. This would break the theoretical guarantee that the χ^2 term stabilizes learning by bounding the learned reward. One could apply Term 3 to the unshifted $\hat{r}$, but this creates an asymmetry: Term 1 operates on the shifted reward while Term 3 operates on the original, making the overall objective internally inconsistent.

Furthermore, the Bellman shift couples the social prior directly with the learned dynamics inside the same objective term. The magnitude of $\lambda \cdot R_{\text{total}}$ must be calibrated relative to the typical scale of $Q(s, a)$ and $\gamma V(s')$. If the ratio $|\lambda \cdot R_{\text{total}}|/|Q|$ is too large, the R_{total} signal dominates the Bellman equation and the Q-function ignores the demonstration data, producing a policy that satisfies the social prior but lacks physical plausibility. If the ratio is too small, the shift has negligible effect. This creates a narrow effective range for λ that depends on the initially unknown scale of the Q-values, making hyperparameter tuning fragile.

B.2. Demonstration Reweighting

The second strategy uses the data distribution rather than the objective function. Each expert transition i is assigned an importance weight proportional to its R_{total} alignment:

$$w_i = \frac{\exp(R_{\text{total}}^{(i)}/\tau_w)}{\sum_{j=1}^{N} \exp(R_{\text{total}}^{(j)}/\tau_w)}, \tag{B.2}$$

where $\tau_w > 0$ is a temperature parameter controlling the sharpness of the weighting. During training, expert mini-batches are sampled according to $\{w_i\}$ rather than uniformly. Transitions with higher R_{total}, those more consistent with the target social profile, are seen more frequently.

While conceptually straightforward, this approach has a fundamental limitation in the fact that it provides no gradient signal about what the recovered reward should look like. The IQ-Learn loss function remains unchanged, only the distribution of transitions it operates on is modified. This means the Q-network never receives an explicit signal that the reward it recovers should resemble R_{total}. The reward is shaped only indirectly, through the statistical composition of the training batches.

This indirection leads to information loss. Consider a transition from an egoistic driver who executes a skilled lane change with high r_{env} but low r_{social}. Under aggressive reweighting (low τ_w), this transition is effectively discarded despite containing valuable information about lane-change dynamics that would benefit any policy, regardless of social alignment. The physical driving competence and the social intent are conflated at the data level, with no mechanism to separate them.

Additionally, heavy reweighting concentrates probability mass on a small subset of transitions, reducing the effective sample size. This ultimately increases gradient variance and slows convergence.

C

IQ-Learn Hyperparameters

Table C.1: Hyperparameter summary. This table summarizes the key hyperparameters for both environments.

Parameter	**Gridworld**	**Intersection**
Hidden layers	[64, 64]	[256, 256]
Learning rate	3×10^{-4}	3×10^{-4}
Batch size	64	256
γ	0.99	0.99
τ (temperature)	1.0	1.0
τ_{target} (soft update)	0.005	0.005
Divergence	χ^2 ($\alpha = 0.5$)	χ^2 ($\alpha = 0.5$)
Loss type	value	value
Regularization mode	reward_reg	reward_reg
Cumulative returns	Yes ($\gamma_{\text{svo}} = 0.99$)	Yes ($\gamma_{\text{svo}} = 0.99$)

D

Lambda Sweep and Convergence Analysis

To understand how the regularization strength λ controls the transition from baseline to target behaviour, we conduct a sweep over $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0\}$ with the egoistic target angle ($\alpha_{\text{target}} = 0°$).

Figure D.1 shows the evolution of $\hat{\alpha}$ during training for each λ value. All runs begin near the baseline angle ($\sim 152°$) and, for sufficiently large λ, shift toward the target angle (-42°). This behavioural shift exhibits a clear phase transition: values of $\lambda \leq 0.5$ fail to overcome the arbitrary mode collapse, so the recovered angle remains near the baseline throughout training (marked with red crosses in Figure D.1, right panel). At $\lambda \approx 0.7$, adaptation begins but is slow and variable across seeds. For $\lambda \geq 1.0$, alignment is reliable and fast, with the agent adopting the target behaviour within the first few evaluation checkpoints.

Figure D.1 (right panel) quantifies this relationship: the number of training updates required to reach $|\hat{\alpha} - \alpha_{\text{target}}| < 20°$drops sharply between $\lambda = 0.7$ and $\lambda = 1.5$, then plateaus. Beyond $\lambda = 1.5$, increasing the regularization strength provides no further benefit in the rate of behavioural adaptation, suggesting that the regularizer has steered the policy away from mode-averaged local optimum.

Finally, Figure D.2 presents six metrics as functions of λ. The empirical recovered angle shows a sharp transition from 152°to -42°around $\lambda = 0.7$, with all runs adopting the target behaviour by $\lambda = 1.0$. The mean Σr_{self} mirrors this transition: it jumps from approximately -5 to approximately +2 as λ crosses the critical threshold. Conversely, Σr_{social} drops from +3 to -2, confirming that the egoistic agent achieves high self-reward at the expense of social impact.

The success rate undergoes a corresponding jump from below 40% to nearly 100%, indicating that the regularizer not only shifts the social character of the policy but also restores the physical driving competence that the baseline had lost. Mean reward and episode length follow consistent patterns: reward increases sharply as the agent adapts to completing the intersection traversal, and episode length decreases as the agent actively drives through rather than waiting indefinitely.

These results confirm a key prediction from the methodology. The reward regularization approach produces monotonic and predictable behavioural shifts as λ varies. Below the critical threshold, the method naturally defaults to the unregularized IQ-Learn baseline. Above it, all behavioural metrics align rapidly. There is no narrow band of policy instability and the behavioural transition is sharp but unidirectional.

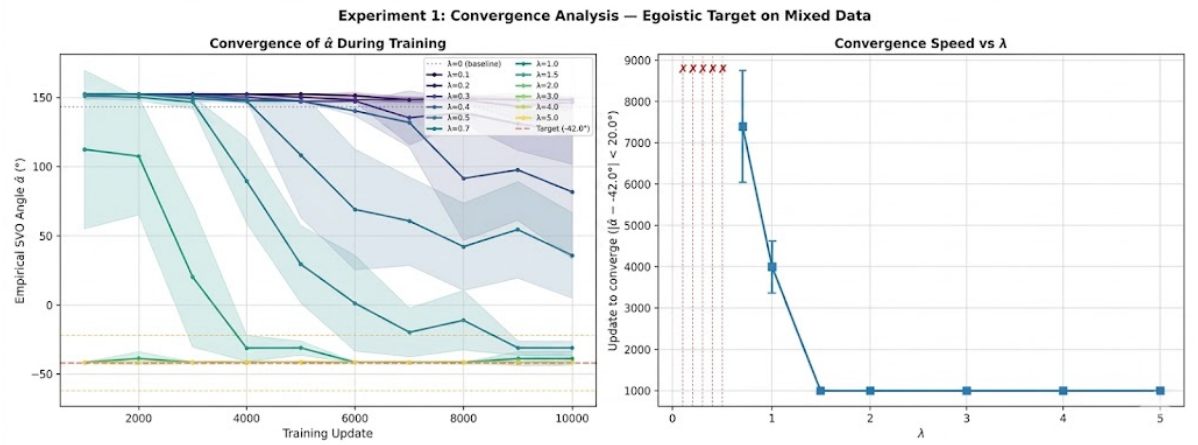


Figure D.1: Adaptation of the recovered social angle across regularization strengths. The left panel illustrates the evolution of the empirical angle over the training process for various λ values. The right panel plots the number of training updates required to approximate the egoistic target angle against the applied λ value.

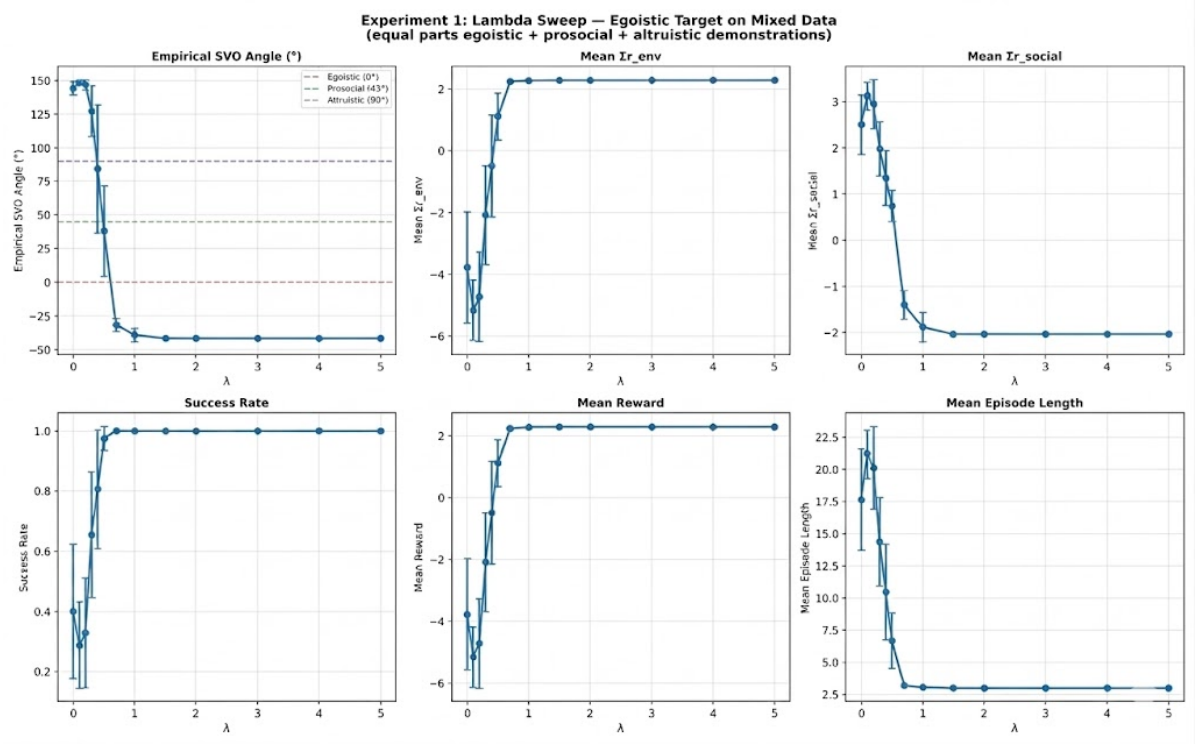


Figure D.2: Behavioural metrics as a function of regularization strength. This figure presents the results of a parameter sweep across λ for the egoistic target profile. The six subplots detail the final empirical angle, mean individual reward, mean social impact reward, success rate, mean total reward, and mean episode length evaluated across the tested λ thresholds.

D.1. Lambda Sensitivity and Practical Tuning

The gridworld λ sweep provides empirical evidence for the theoretical claim made in Section 4.3.1, that the reward regularization mode produces monotonic and predictable behaviour as λ varies. This behaviour is precisely the empirical signature predicted by Tikhonov regularization theory. The MSE penalty $\|\hat{r} - \lambda \cdot R_{\text{total}}\|^2$ acts as a quadratic well centered on the social target. As λ increases, the social prior exerts progressively stronger influence on the recovered reward, with the IQ-Learn objective continuing to shape the reward on expert-visited states. As λ increases from zero, the regularizer initially does not have sufficient strength to overcome the IQ-Learn objective's mode averaging tendency. Beyond a critical

region, the social prior begins to visibly redirect the recovered policy, and the empirical recovered angle progresses smoothly toward the target as λ continues to grow. Once λ grows large enough, the agent reliably aligns with the target behaviour. Pushing the regularization strength even further gives stable plateauing results, proving that there are no fragile tuning zones and no unpredictable behavioural swings at extremely high λ values.

This contrasts with the predicted behaviour of the Bellman shift approach (Section B.1), where λ must be calibrated relative to the evolving Q-scale, creating a fragile tuning problem. In the reward regularization mode, the critical threshold is a property of the environment and dataset, not of the learned Q-values. Once the designer determines that λ is above this threshold, which can be verified by monitoring the recovered $\hat{\alpha}$ during training, the exact value is not critical. In practice, this means the method requires a coarse search (is λ large enough?) rather than a fine-grained calibration (is λ in the right narrow band?), substantially reducing the tuning burden. Ultimately, these results empirically validate the choice of reward regularization as the integration strategy (Contribution 1), confirming that the Tikhonov formulation provides the predictable, robust control over the recovered reward that alternative strategies lack.

E

AI Usage

This appendix discloses the usage of generative AI during the research and development process of this thesis, in accordance with the TU Delft guidelines. Specifically, Gemini and Claude were used in the research workflow. These tools were used to assist in refining methodological approaches, finding literature references, and debugging the underlying code. Additionally, they provided assistance for brainstorming ideas and improving the grammatical fluency and LaTeX formatting of the thesis. All outputs generated by Gemini and Claude were inspected for technical accuracy and reliability before including them in the thesis. Finally, no sensitive, proprietary, or confidential data was exposed to these models at any stage of the project.