Towards Safety Aware AI Agents

Thomas Steinecker*, Thorsten Luettel* and Mirko Maehlisch*

Abstract:

Safety is the most critical aspect to address for the real-world deployment of robotic platforms, such as autonomous driving systems. While learning-based approaches like reinforcement learning have gained popularity for managing real-world complexity, they often lack transparency and safety awareness. In this paper, we aim to advance the development of safety-aware AI agents by presenting a framework for estimating collision probability distributions, which can be integrated into the decision-making process of reinforcement learning agents. To this end, we provide a thorough definition and motivation for incorporating safety awareness, highlighting its importance for reliable and interpretable decision-making. Finally, we demonstrate how these collision probabilities can be effectively integrated into decision-making by incorporating them into a value function, enabling safety-aware reinforcement learning.

Keywords: collision avoidance, reinforcement learning, safety awareness, temporal difference learning

1 Introduction

In safety-critical systems, interpretability is essential [1], which is why interest in explainable AI (XAI) has significantly grown in recent years [2]. Efforts to enhance decision transparency have been made across various robotic domains [3], driven in part by concerns over the lack of trust in deep learning-driven approaches [4], particularly in autonomous driving [5, 6]. A very promising approach for machine learning based decision making is reinforcement learning (RL), due to its recent successes in many areas [7, 8]. Even though safety is typically incorporated by assigning negative rewards for collisions [9] or proximity to other objects [10, 11], explicit risk assessment is absent. The overall quality measure for a state is represented as a single value from the value function, which is primarily used during training and has no direct effect during deployment. Whereas some approaches exist that try to include risk assessment by extending the framework with other tools, such as prediction modules [10], there is still no holistic approach for safety-aware RL.

To address this issue, we propose modifying the value function in reinforcement learning to reveal situational risk via collision probability distribution estimation [12], which can be interpreted as a part of the value function decomposition [13]. The total value can then be used for decision-making through standard RL approaches [14]. This provides two key benefits: (1) the collision probability distribution serves as a transparent,

*This research paper [project MORE] is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

All authors are with the Institute for Autonomous Systems Technology (TAS), Department of Aerospace Engineering, University of the Bundeswehr Munich, Neubiberg, Germany. Contact author email: thomas.steinecker@unibw.de

The code is available at `https://github.com/UniBwTAS/CollisionPro`

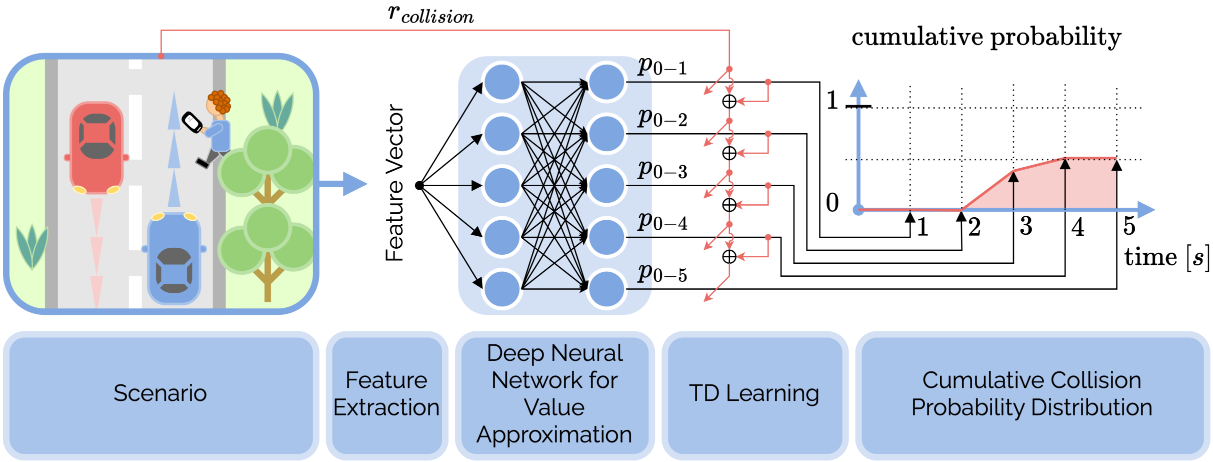

Figure 1: An overview of collision probability distribution estimation as described in [12]. Given a scenario where a feature vector is extracted (which could also be a raw camera image), a neural network predicts cumulative collision probability values over a specified time horizon (in this case, 5 seconds), with equidistant time intervals between all predicted probabilities (in this case, 1 second). The learning procedure utilizes a bootstrapping mechanism, meaning that collision probability values for shorter time horizons can learn from those for longer time horizons.

interpretable metric during both training and deployment, and (2) the learned collision probabilities can offer valuable feedback to the agent. An overview of our collision probability distribution estimation is given in Fig. 1.

Our approach does not guarantee safety as formal methods do [15, 16], instead it embraces the fact, that collisions can always occur in reality due to uncertainty arising from model errors or perception limitations. We argue that this approach offers several advantages, including more efficient training, as risk estimates can be leveraged for scenario generation and safety comparisons between different approaches.

Building on our work in [12], we frame our approach in the context of safety awareness, shift from policy evaluation to policy improvement, and demonstrate its applicability to long-term (5-second) predictions. These predictions can aid reinforcement learning algorithms in high-level decisions, such as merging into adjacent lanes. While we present the overall framework, our primary focus remains on policy evaluation—specifically, estimating collision probability distributions.

In the following section, we introduce the concept of safety awareness and its relevance to our approach. Sec. 3 provides an overview of related work on collision probability estimation. In Sec. 4, we present the collision probability distribution estimation framework, initially focusing on policy evaluation and later extending it to policy improvement. Sec. 5 presents an evaluation of the policy evaluation in an autonomous driving simulator (CARLA Simulator [17]). In Sec. 6, we provide a brief discussion of our findings and outline future research directions. Finally, we conclude the paper in Sec. 7.

2 Safety Awareness in Decision-Making

Safety itself can be viewed from multiple perspectives, including mechanical limitations, environmental constraints, or the probability of collisions. In this work, we focus specifically on collisions, as they represent a crucial and directly measurable aspect of safety in autonomous systems.

2.1 What is safety awareness?

Safety awareness refers to an agent's ability to assess risks based on its own behavior and understanding of its environment. Inspired by the general concept of awareness—defined as "*knowledge that something exists, or understanding of a situation or subject at the present time based on information or experience*" (Cambridge Dictionary)—safety awareness encompasses an agent's ability to recognize and interpret potential risks based on its interactions/experiences with the world. Unlike predefined or externally imposed safety metrics, safety awareness emerges as a result of an agent's behavior within a given environment, making risk not an independent variable but a direct consequence of the agent's decisions and actions. Many existing approaches define risk or safety independently of the agent's behavior; however, this neglects the fundamental fact that risk is inherently behavior-dependent.

We distinguish between **implicit** and **explicit** safety awareness. **Explicit safety awareness** refers to methods where the risk assessment is explicitly represented and communicated, allowing external observers or other components of the system to interpret the agent's safety evaluation. In contrast, **implicit safety awareness** occurs when the agent internally accounts for safety in its decision-making process without explicitly exposing a risk measure. Vanilla reinforcement learning strategies typically exhibit implicit safety awareness, as the value function is a composition of many different reward signals, such as safety, energy efficiency, and performance.

Interpretability is an essential aspect of safety awareness, as it enables transparent risk assessment and facilitates meaningful comparisons across different policies, states or actions. Traditional reinforcement learning approaches often lack interpretability because value functions aggregate multiple reward signals, making it difficult to isolate safety-related information. Additionally, the use of a discount factor introduces temporal scaling effects, further obscuring the direct assessment of risk.

In contrast to formal safety guarantee methods, which aim to ensure absolute safety through strict constraints and verification techniques, **safety awareness focuses on estimating and assessing risk rather than eliminating it.** This distinction makes safety awareness particularly relevant for model-free approaches, where no structured abstraction of the environment exists, rendering traditional formal methods inapplicable. Instead of relying on strict verification for absolute safety guarantees, safety-aware agents estimate and assess risk, allowing them to adapt to uncertain environments.

The necessity of safety awareness arises fundamentally from uncertainty. In dynamic and unpredictable environments, agents must continuously evaluate potential hazards rather than rely on static safety assurances. Safety awareness is not merely about ensuring safety but about developing a deeper understanding of the conditions that influence it. Importantly, this perspective acknowledges that absolute safety may not always be achievable but can still be meaningfully assessed and managed.

2.2 Collision Probability Distribution in the Context of Safety Awareness

In this paper we propose a collision probability distribution estimation based on temporal differences which is briefly explained in Sec. 4.1 and evaluated in Sec. 5, as part of the decision-making process of an RL agent, which is explained in Sec. 4.2. Our approach provides explicit, interpretable safety awareness and in the following we want to account for its beneficial properties:

- **Interpretability**
  The estimation of collision probability differs significantly from the traditional value function approach used in reinforcement learning (RL). Unlike the value function (see Sec. 2.1), which requires a discount factor, the estimated collision probability offers interpretability.

- **Comparability**
  One of the challenges of using the value function is its sensitivity to different discount factors, reward structures, and parameter settings. This sensitivity makes it difficult to compare results across varying reward configurations. In contrast, explicitly estimating collision probabilities provides a more consistent framework for evaluating safety. It enables a quantitative comparison between different policies with respect to the overall performance, but also for specific scenes/states.

- **Temporal Information**
  The proposed approach enables the estimation of time windows in which collisions are likely to occur. This temporal information can be utilized to communicate imminent risks to external systems or passengers, allowing for proactive warnings and timely intervention.

- **Targeted Training and Scenario Generation**
  Transparency in safety evaluation facilitates the identification of hazardous scenarios, which can then be used to refine training and improve agent performance. Scenarios are classified as dangerous for two primary reasons: (1) the agent's suboptimal actions increase the likelihood of a collision, or (2) the scenario inherently leads to a higher risk of collision despite optimal agent behavior. By recognizing these distinctions, targeted scenarios can be generated to address specific weaknesses in the agent's decision-making process, ultimately enhancing its robustness in critical situations.

- **Isolated Safety Evaluation**
  A limitation of the value function in traditional RL frameworks is the aggregation of values across multiple criteria, which precludes an isolated assessment of safety. By focusing explicitly on collision probabilities, the proposed method enables a decoupled evaluation of safety-related aspects. This isolation enhances interpretability and provides more actionable insights into the agent's performance in safety-critical contexts.

- **Avoiding Redundancy**
  Many RL approaches incorporate risk assessment through additional methods, such as state prediction neural networks [10]. However, our approach eliminates the need for such auxiliary models by directly estimating collision probabilities within the decision-making process.

3 Related Work

Many widely recognized approaches in autonomous driving do not explicitly incorporate collision probabilities. Notable examples include the rule-based Responsibility-Sensitive Safety (RSS) framework by MobileEye [18] and the Safety Force Field (SFF) by Nvidia [19]. These methods rely on predefined behavioral assumptions about other traffic participants and object dynamics. While they have been successfully applied and remain subjects of ongoing research, they inherently overlook critical aspects such as the aggressiveness of individual drivers or regional variations in driving behavior. Essentially, these approaches operate under the assumption that as long as predefined safety rules are followed, the probability of collision remains zero or negligible.

In contrast, some approaches explicitly compute collision probabilities. For instance, [16] employs a Monte Carlo (MC) method to estimate a single collision probability involving both static and dynamic objects, which is then used to determine appropriate speed limits. Similarly, [15] uses a more sophisticated analytic approach, leveraging stochastic reachable sets to compute collision probabilities.

The concept of collision probability distributions (CPDs) is not novel and has been explored within classical methodologies. In [20], Monte Carlo simulations are applied to a Gaussian distribution of trajectories for both the ego vehicle and surrounding traffic participants, allowing the computation of a time-dependent CPD. A more efficient alternative is presented in [21], where dynamic objects are approximated using octagonal bounding regions, leading to an analytical solution with improved computational performance. These methods perform CPD estimation entirely online, whereas our approach leverages deep learning. In our framework, CPDs are learned during training, and only inference is performed online, enabling real-time application with pre-trained experience.

Unlike classical methods, machine learning-based approaches are not necessarily constrained by the same simplifying assumptions. For instance, [22] proposes a deep predictive model that estimates both the mean and variance of collision probability using variational inference [23]. Other approaches rely on scene graphs to model interactions. In [5], a scene graph captures the topological relationships between vehicles, while [24] extends this concept to pedestrian interactions to estimate collision probabilities.

Despite these advancements, none of the aforementioned approaches compute a full CPD within a learning-based framework, nor do they integrate CPDs into reinforcement learning framework. Our method addresses this gap by providing a deep-learning-based CPD estimation technique that can be seamlessly incorporated into reinforcement learning agents. Tab. 1 summarizes the key differences between our approach and prior work.

Table 1: Classification of collision probability approaches

	Single probability value	Probability distribution
Classical	[15, 16]	[20, 21]
Learning-based	[5, 6, 22]	**Ours**

4 Methods

4.1 Collision Probability Distribution Estimation

The main idea of our approach is to use a **fixed, finite horizon** ($T_H = \Delta t \cdot N_H$), where Δt represents the time discretization step, and to assume no discount factor ($\gamma = 1$). This formulation can be derived similarly to the Bellman equation:

$$\begin{aligned} V_{t\to t+N_H}(s) &= \mathbb{E}[R_{t+1} + R_{t+2} + \ldots + R_{t+N_H} \mid S_0 = s] \\ &= \mathbb{E}[R_{t+1} + G_{t+1\to t+N_H} \mid S_0 = s] \\ &= \mathbb{E}[R_{t+1} + V_{t\to t+N_H-1}(s_{t+1}) \mid S_0 = s] \end{aligned} \tag{1}$$

Here, $V_{t\to t+i}(s)$ is the finite-horizon value function for state s with a horizon of $\Delta t \cdot i$, and $G_{t+k\to t+l}$ is the return from time $t+k$ to $t+l$. It becomes clear that the bootstrapping mechanism works by using value functions with progressively smaller horizons. This requires a **distribution of value functions**, but simultaneously provides richer outputs by **incorporating temporal information**. By introducing multiple value functions:

$$V_{t\to t+i}, \quad \text{for} \quad i = 1, \ldots, N_H \tag{2}$$

we can apply temporal difference learning, which can be further extended to TD(λ) [12, 14], providing a powerful tool for value function learning.

Next, we assign a reward of -1 for collisions and 0 otherwise, leading to the following relationship:

$$p_{collision,t\to t+i}(s) = -V_{t\to t+i}(s) \tag{3}$$

where $p_{collision,t\to t+i}(s)$ is the collision probability within the time interval $[t, t+i]$. This arises from the fact that the probability of an event occurring within a given time horizon is the number of times the event occurs from a given state divided by the total number of times that state is visited, which directly corresponds to the definition in Eq. (1). For more details on the framework, we refer the reader to [12].

4.2 Safety Awareness

In the previous section, we formulated a method to calculate the collision probability distribution for a finite horizon. This distribution can be represented as a vector of values:

$$\mathbf{V} = \begin{bmatrix} V_{t\to t+1} & V_{t\to t+2} & \ldots & V_{t\to t+N_H} \end{bmatrix}^T \tag{4}$$

However, in RL, a scalar is typically used to assess the value of a state. Therefore, we need an aggregation function to derive a single value from Eq. (4):

$$V_{collision} = f_{aggregate}(\mathbf{V})\colon \mathbf{V} \in \mathbb{R}^{N_H} \to V_{collision} \in \mathbb{R} \tag{5}$$

An overview of this approach is given in Fig. 2. The aggregation function could be a simple min-operator, meaning that only the collision probability for the entire time horizon is considered, while the distribution itself is ignored. Note that the min-operator must be considered because of the relationship given in Eq. (3). However, we argue that a more sophisticated function that takes the entire distribution into account may be more beneficial. Potential collision events further in the future could provide valuable time for additional actions, such as warning the vehicle occupants or mitigating the severity of the collision. This idea is illustrated in Fig. 3.

Once the collision probability distribution has been aggregated, as specified in Eq. (5), we can use the value function decomposition approach to obtain a single value function [13].

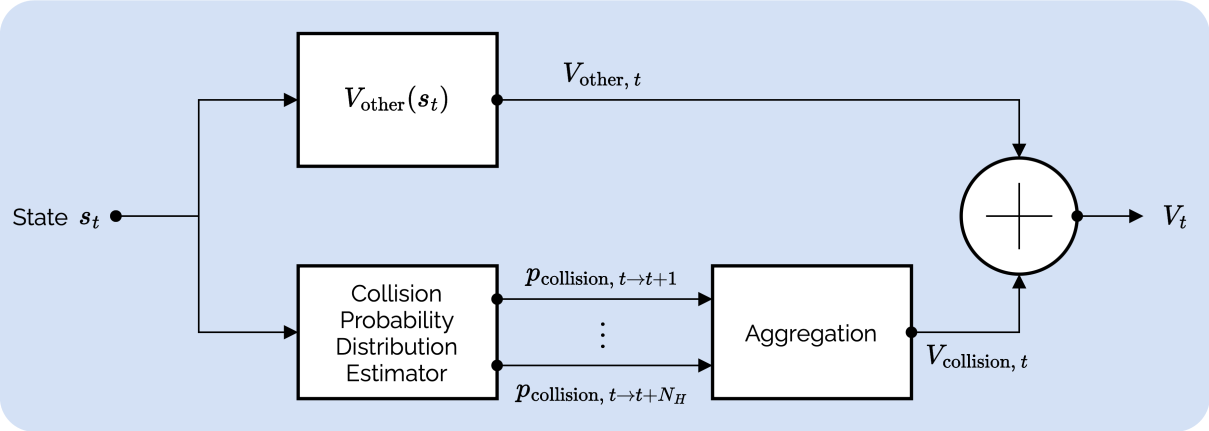
State s_t
$V_{\text{other}}(s_t)$
$V_{\text{other},\,t}$
Collision Probability Distribution Estimator
$p_{\text{collision},\,t\to t+1}$
⋮
$p_{\text{collision},\,t\to t+N_H}$
Aggregation
$V_{\text{collision},\,t}$
V_t

Figure 2: A proposal to integrate the collision probability distribution estimation into the value function.

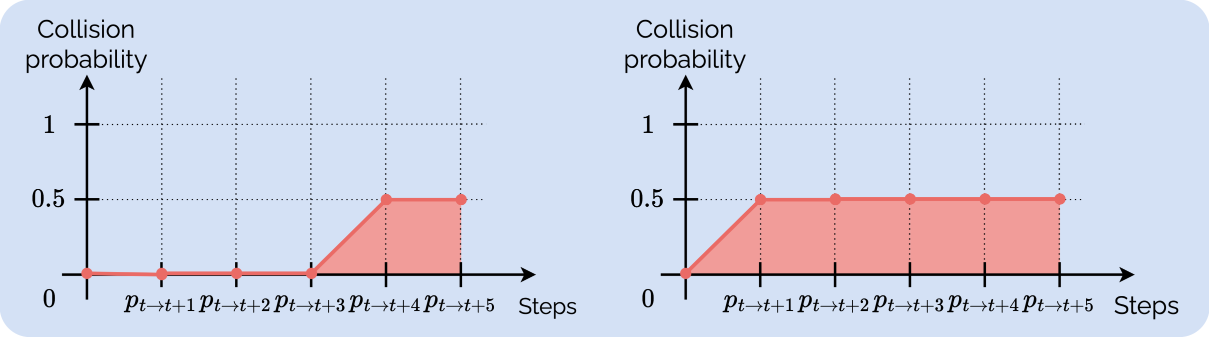

Figure 3: Both subfigures display collision probability distributions. While the overall collision probability for the entire horizon ($N_H = 5$) is the same ($p_{t\to t+N_H} = 0.5$) in both cases, the left subfigure indicates that the collision event is most likely to occur 3 steps later. This delay is more advantageous, as it allows time for additional actions, such as switching strategies (e.g., safe braking) or issuing alerts.

5 Evaluation

Our evaluation focuses on the prediction capabilities of the collision probability distribution estimation, using a long-term horizon of 5 seconds with a time step of 0.1 seconds. We used the CARLA Simulator [17] by spawning a fleet of 30 vehicles and pedestrians, all autonomously controlled by the Traffic Manager. The ego vehicle also operated autonomously, following a simple lane-keeping strategy.

The evaluation was conducted over 2000 episodes, each with a maximum length of 3000 steps, during which 744 episodes ended in a collision. The value function approximator was a deep neural network with a CNN backbone, which received stacked semantic bird's-eye views as input [25]. Further details on the architecture are available in [12].

In Fig. 4, the collision probability distributions over the final 5 seconds before a collision are visualized. The results demonstrate that our framework effectively learns collision probabilities, with predictions becoming increasingly confident as the ego vehicle approaches the collision, as indicated by the steadily rising probability values. However, some collision characteristics in the figure suggest that certain events were detected relatively late, likely due to a limited feature space lacking sufficient evidence or an insuf-

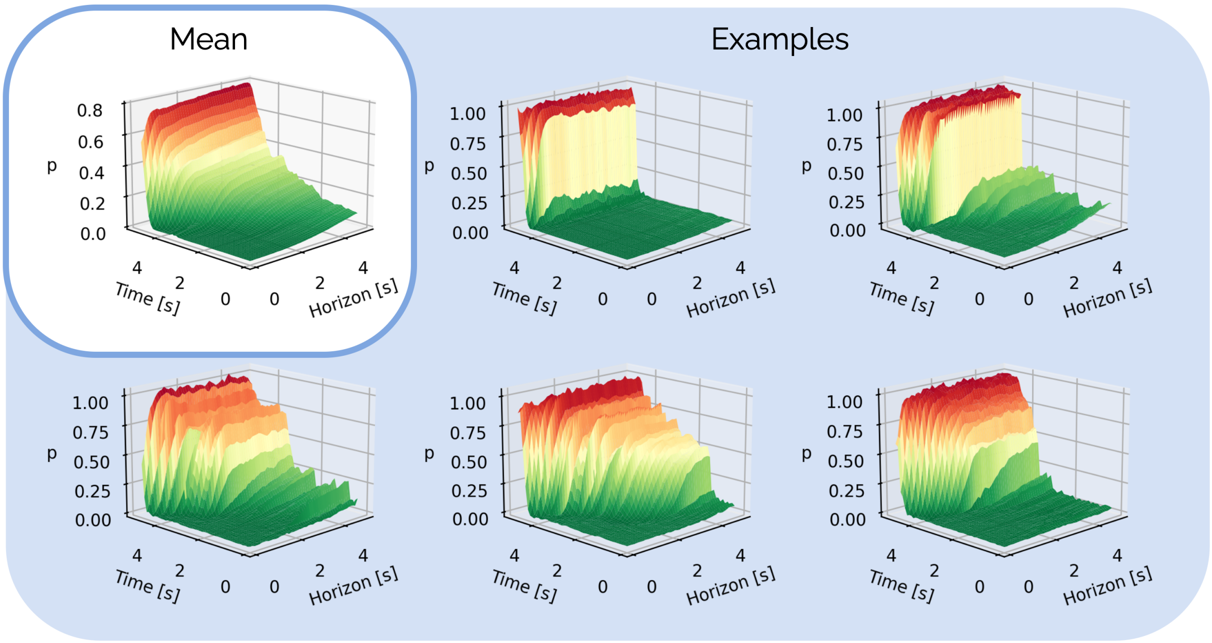

Figure 4: Collision characteristics in the testing environment within the CARLA Simulator. The visualization illustrates the evolution of collision probabilities over time, leading up to a collision event at $t = 5.0$ seconds. The top left subfigure depicts the mean collision probability over 50 collision events, while the remaining subfigures present individual examples.

ficient training dataset. Nonetheless, the mean collision characteristics shown in the top left subfigure of Fig. 4 reflect a robust risk assessment.

To the best of our knowledge, this is the first work to introduce a generic, model-free framework for estimating temporal collision probability distributions. Consequently, we did not conduct a direct comparative evaluation with existing approaches.

6 Discussion and Future Work

The approach presented in Sec. 4.1 establishes a foundational framework for safety-aware agents by integrating situational risk assessment into their decision-making processes. In our evaluation, we considered a 5-second time horizon for collision probability estimation, which is sufficient for making high-level driving decisions, such as merging into adjacent lane.

While the proposed approach demonstrates promising results, it currently relies on collision experience, which is only feasible within simulated environments such as 3D physics engines or learned world models [26]. Consequently, bridging the sim-to-real gap remains a crucial challenge, as transferring learned safety-aware behaviors to real-world driving scenarios requires addressing domain discrepancies–an ongoing research topic in the field.

Another important limitation is the scalability of neural networks in handling rare collision events. In real-world autonomous driving, collisions are exceedingly infrequent. This raises concerns about data sparsity, as the network may struggle to learn meaningful patterns when collision probabilities are extremely small. A potential strategy to improve learning efficiency under these conditions could be to apply an alternative output scaling, such as a logarithmic transformation, to better capture low-probability events. Further investigation is needed to explore how different output representations impact model performance.

This work introduces a framework for integrating collision probability estimation into standard RL pipelines. However, our evaluation has so far been limited to a single fixed policy, focusing only on policy evaluation. Future work will extend this framework to iteratively improve policies through reinforcement learning, enabling an agent to actively optimize its behavior based on safety-aware collision probability estimates. Exploring this direction will be crucial for realizing fully autonomous, risk-aware decision-making in complex environments.

7 Conclusion

This paper presents a safety-aware AI framework that estimates collision probability distributions and seamlessly integrates them into the reinforcement learning (RL) framework. Unlike existing RL approaches that rely on indirect safety measures, our method provides an explicit and interpretable risk assessment mechanism. So far, we have demonstrated successful policy evaluation, and in future work, we will investigate how our framework can be leveraged for policy improvement in safety-critical scenarios.

References

[1] T. Steinecker, A. Kurdas, N. Mansfeld, M. Hamad, R. J. Kirschner, S. Abdolshah, and S. Haddadin, “Mean reflected mass: A physically interpretable metric for safety assessment and posture optimization in human-robot interaction,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 11 209–11 215.

[2] F. Sado, C. K. Loo, W. S. Liew, M. Kerzel, and S. Wermter, “Explainable goal-driven agents and robots-a comprehensive review,” *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–41, 2023.

[3] C. Glanois, P. Weng, M. Zimmer, D. Li, T. Yang, J. Hao, and W. Liu, “A survey on interpretable reinforcement learning,” *Machine Learning*, pp. 1–44, 2024.

[4] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine *et al.*, “Llama guard: Llm-based input-output safeguard for human-ai conversations,” *arXiv preprint arXiv:2312.06674*, 2023.

[5] A. V. Malawade, S.-Y. Yu, B. Hsu, D. Muthirayan, P. P. Khargonekar, and M. A. Al Faruque, “Spatiotemporal scene-graph embedding for autonomous vehicle collision prediction,” *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9379–9388, 2022.

[6] E. Candela, O. Doustaly, L. Parada, F. Feng, Y. Demiris, and P. Angeloudis, “Risk-aware controller for autonomous vehicles using model-based collision prediction and reinforcement learning,” *Artificial Intelligence*, vol. 320, p. 103923, 2023.

[7] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, “Mastering the game of go without human knowledge,” *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[8] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza, "Champion-level drone racing using deep reinforcement learning," *Nature*, vol. 620, no. 7976, pp. 982–987, 2023.

[9] G. Chen, Y. Zhang, and X. Li, "Attention-based highway safety planner for autonomous driving via deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, 2023.

[10] A. Baheri, S. Nageshrao, H. E. Tseng, I. Kolmanovsky, A. Girard, and D. Filev, "Deep reinforcement learning with enhanced safety for autonomous highway driving," in *2020 IEEE Intelligent Vehicles Symposium (IV).* IEEE, pp. 1550–1555.

[11] K. Ahmic, J. Ultsch, J. Brembeck, and C. Winter, "Reinforcement learning-based path following control with dynamics randomization for parametric uncertainties in autonomous driving," *Applied Sciences*, vol. 13, no. 6, p. 3456, 2023.

[12] T. Steinecker, T. Luettel, and M. Maehlisch, "Collision probability distribution estimation via temporal difference learning," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC).*

[13] J. MacGlashan, E. Archer, A. Devlic, T. Seno, C. Sherstan, P. Wurman, and P. Stone, "Value function decomposition for iterative design of reinforcement learning agents," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 001–12 013, 2022.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* MIT press, 2018.

[15] M. Althoff, O. Stursberg, and M. Buss, "Model-based probabilistic collision detection in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 2, pp. 299–310, 2009.

[16] V. Kibalov and O. Shipitko, "Safe speed control and collision probability estimation under ego-pose uncertainty for autonomous vehicle," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC).* IEEE, pp. 1–6.

[17] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Conference on robot learning.* PMLR, 2017, pp. 1–16.

[18] S. Shalev-Shwartz, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374*, 2017.

[19] D. Nistér, H.-L. Lee, J. Ng, and Y. Wang, "The safety force field," *NVIDIA White Paper*, vol. 15, 2019.

[20] S. Annell, A. Gratner, and L. Svensson, "Probabilistic collision estimation system for autonomous vehicles," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC).* IEEE, pp. 473–478.

[21] A. Philipp and D. Goehring, "Analytic collision risk calculation for autonomous vehicle navigation," in *2019 International Conference on Robotics and Automation (ICRA).* IEEE, pp. 1744–1750.

[22] M. Strickland, G. Fainekos, and H. B. Amor, "Deep predictive models for collision risk assessment in autonomous driving," in *2018 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, pp. 4685–4692.

[23] Y. Gal *et al.*, "Uncertainty in deep learning," 2016.

[24] X. Liu, Y. Zhou, and C. Gou, "Learning from interaction-enhanced scene graph for pedestrian collision risk assessment," *IEEE Transactions on Intelligent Vehicles*, 2023.

[25] J. Chen, B. Yuan, and M. Tomizuka, "Deep imitation learning for autonomous driving in generic urban scenarios with enhanced safety," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 2884–2890.

[26] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.