Risk-based Socially-Compliant Behavior Planning for Autonomous Driving

Yiwei Lyu[1], Wenhao Luo[2] and John M. Dolan[1]

***Abstract*— In this study, we introduce an innovative risk-aware behavior planning framework designed for autonomous driving, with the aim of fostering socially compliant vehicle behavior in diverse mixed-traffic highway scenarios. Our objective is to empower autonomous vehicles to exhibit behavior that aligns with societal norms, thus enhancing their acceptability among human drivers. We expand the scope of Control Barrier Function-inspired risk assessment to encompass a heterogeneous spectrum of road participants, allowing us to explicitly model varying degrees of social influences between different classes of vehicles. We also present a mathematical condition for accountability tracing, enabling the identification of responsible entities in situations where risks surge. Drawing inspiration from Isaac Asimov's "Three Laws of Robotics," we establish social compliance conditions grounded in our unique risk concept, which seamlessly integrates with a wide range of existing safety-critical controllers, regardless of their type or design. By incorporating these conditions, which encode societal expectations, into existing safe controllers, we demonstrate that autonomous vehicles can exhibit context-aware behavior without compromising the safety guarantees provided by existing controllers. This approach effectively excludes behaviors that may be safe but do not align with human intuition while guaranteeing the least interference with the existing controller.**

I. Introduction

There have been remarkable advancements in autonomous driving technology, leading to the deployment of self-driving cars on real-world streets. These autonomous vehicles now share the road with human drivers, marking a significant milestone in the progression of this technology. It is increasingly clear that the coexistence of autonomous vehicles and human drivers is not merely a concept but a practical reality, and it is anticipated that large-scale mixed-traffic scenarios will become increasingly common in the near future.

To foster a harmonious coexistence between self-driving cars and human-driven vehicles, prioritizing safety is paramount. Various techniques and tools [1], [2], [3], [4] have been explored to enhance the vehicle's ability to avoid possible collisions. However, safety considerations should not be confined solely to physical state configurations, such as avoiding collisions with human drivers. They should also extend to encompass the psychological aspect of human perception and trust [5], [6], [7], [8]. Ensuring that interactions with self-driving cars instill a sense of safety and confidence is crucial. Therefore, it is equally imperative to explore innovative approaches that empower autonomous vehicles to exhibit socially compliant behavior. By doing so, these vehicles can align their actions with human expectations, enhancing their overall acceptability among human drivers.

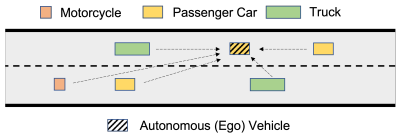

Fig. 1. An example scenario of highway driving with mixed classes of traffic participants driving from left to right. The box with diagonal strips represents the ego vehicle under our control. Other surrounding vehicles are marked in different colors based on their types, pink for motorcycles, yellow for passenger cars, and green for trucks. The dash arrows represent the risk each surrounding vehicle poses on the ego vehicle in the pairwise relationship.

Numerous studies have explored the concept of human-like vehicle control [9], [10], [11], often relying on cost functions that are either manually crafted or learned from data, along with metrics like Root Mean Square Error and Average Displacement Error. These approaches have proven effective in replicating human trajectories from datasets, bolstering repeatability. However, the challenge of adaptability persists, as real-world scenarios can vary widely making it challenging to be adequately captured and represented in training data. Consequently, there is a pressing need for a unified framework capable of generating vehicle behavior that is acceptable to humans across a wide range of scenarios. There have been efforts to integrate traffic rules into control frameworks to enable scenario-aware behavior for self-driving cars [12], [13]. Many of these approaches rely on rule-based methods, where different rules are integrated into the autonomous driving control problem. However, a challenge arises in that distinct rules must be specified for various driving scenarios. Furthermore, it's important to highlight that these objective rules may not entirely capture the subjective nuances that differentiate, for example, the behavior of an autonomous vehicle suddenly merging in front of a heavily loaded truck on the highway, ensuring no collisions occur and all rules respected, from behavior that aligns with typical human expectations.

In this work, we introduce a novel behavior planning framework that relies on risk assessment as its foundational concept. Risk evaluation has been a well-explored topic in the realm of robot control, with some approaches incorporating it into the objective function to minimize the risks faced

*This work was supported in part by the Qualcomm Innovation Fellowship, in part by the Faculty Research Grant award at UNC Charlotte, and in part by the U.S. National Science Foundation under Grant CNS-2312465.

[1] The authors are with Carnegie Mellon University, Pittsburgh, USA. Email: yiweilyu,jdolan@andrew.cmu.edu

[2]The author is with the Department of Computer Science, University of North Carolina at Charlotte, Charlotte NC 28223, USA. Email: wenhao.luo@uncc.edu.

by agents in their environments [14], [15]. However, these approaches often lead to unintended overly conservative behavior, hindering expected task performance. Moreover, existing risk evaluation methods tend to assess the influence of limited factors, such as the positions and motion of robots. In our previous work [16], [17], we proposed an innovative model-based risk evaluation tool capable of considering additional dimensions, including a robot's safety radius and behavioral aggressiveness. Nevertheless, this risk evaluation tool was primarily designed for homogeneous mobile robots operating in open spaces, rather than the complex scenarios encountered in autonomous driving, where various types of vehicles coexist. We argue that risk assessment should inherently incorporate the heterogeneity of different traffic participants. For instance, the risk posed by a fully loaded truck should not be equated with that of a small passenger vehicle, even if they share the same physical states. Acknowledging and addressing these distinctions is paramount for improving risk evaluation, particularly when tailoring it specifically for autonomous driving systems.

Our **main contributions** are: **1)** We extend the CBF-based risk evaluation to encompass heterogeneous traffic participants, allowing for explicitly modeling the varying degrees of social influence exerted by different vehicle types, a critical consideration in real-world driving scenarios. Based on this extended risk assessment, we formulate the accountability tracing problem in a mathematically quantifiable manner, providing a valuable tool for future policy studies concerning incidents involving self-driving cars. **2)** Drawing inspiration from Isaac Asimov's "Three Laws of Robotics," we derive conditions related to the notion of risk that characterize robot behavior aligning with human instinct and common expectations. These conditions offer broad applicability across various driving scenarios and even in different domains beyond autonomous driving. **3)** By integrating these conditions, which encode social norms, into existing vehicle control problems, we enable autonomous vehicles to exhibit behavior that aligns with typical human expectations while preserving safety. This approach effectively excludes behaviors that may be technically safe but do not align with human intuition while guaranteeing the least interference with the nominal controller.

II. Preliminaries

A. Control Barrier Function

Control Barrier Functions (CBF) [18] are used to define an admissible control space for safety assurance of dynamical systems. One of CBF's important properties is its forward-invariance guarantee of a desired safety set. Consider a nonlinear system in control affine form: $\dot{x} = f(x) + g(x)u$, where $x \in \mathcal{X} \subset \mathbb{R}^n$ and $u \in \mathcal{U} \subset \mathbb{R}^m$ are the system state and control input with f and g assumed to be locally Lipschitz continuous. A desired safety set $\mathcal{H}$ can be denoted by a safety function $h(x)$: $\mathcal{H} = \{x \in \mathbb{R}^n : h(x) \geq 0\}$. Thus the control barrier function for the system to remain in the safety set can be defined as follows [18]:

Definition 1: (Control Barrier Function) Given the aforementioned dynamical system and the set $\mathcal{H}$ with a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$, then h is a control barrier function (CBF) if there exists a class $\mathcal{K}$ function for all $x \in \mathcal{X}$ such that $\sup_{u \in \mathcal{U}} \{\dot{h}(x, u)\} \geq -\kappa\big(h(x)\big)$.
We selected the same class $\mathcal{K}$ function $\kappa(h(x)) = \gamma h(x)$ as in [19], [20], where $\gamma \in \mathbb{R}^{\geq 0}$ is a CBF design parameter controlling system behaviors near the boundary of $h(x) = 0$. Hence, the admissible control space can be redefined as $\mathcal{B}(x) = \{u \in \mathcal{U} : \dot{h}(x, u) + \gamma h(x) \geq 0\ \}$. It is proved in [18] that any controller $u \in \mathcal{B}(x)$ will render the safe state set $\mathcal{H}$ forward-invariant, i.e., if the system starts inside the set $\mathcal{H}$ with $x(t = 0) \in \mathcal{H}$, then it implies $x(t) \in \mathcal{H}$ for all $t > 0$ under controller $u \in \mathcal{B}(x)$.

B. CBF-inspired Risk Evaluation for Pairwise Vehicles

Consider a driving scenario with a total number of vehicles $N \in \mathcal{N}$, in which every vehicle has access to observations of all vehicles' current positions and velocities, but no direct communication is available among vehicles. Similar to [3], [21], [22], we consider the particular choice of pairwise safety function $h_{ij}(x)$ and safety set $\mathcal{H}_{ij}(x) = \{x \in \mathcal{X} : h_{ij}(x) = ||x_i - x_j||^2 - D_{\text{safe}}^2 \geq 0, \forall i \neq j\}$, and admissible control space $\mathcal{B}_{ij}(x) = \{u \in \mathcal{U} : \dot{h}_{ij}(x, u) \geq -\gamma(h_{ij}(x))\}$ for each vehicle pair, where $x_i, x_j \in \mathbb{R}^2$ for $i, j \in \{1, ..., N\}$ are the positions of any pairwise vehicles i and j. We consider single-integrator dynamics $\dot{x} = u$ as in [22] for simplicity, but higher-order dynamics like unicycle dynamics can be achieved using a nonlinear inverse method for velocity mapping [23], [24], [17]. $u = \{u_i, u_j\} \in \mathbb{R}^2$ is the joint control input of this particular vehicle pair, and D_{safe} is the pre-defined safety margin.

Next, to quantify the risk between each pair of vehicles from potential collision, we draw inspirations from CBF and propose the following pairwise safety loss function $L_{ij}(x, u)$:

$$\begin{aligned} L_{ij}(x, u) &= -\dot{h}_{ij}(x, u) - \gamma h_{ij}(x) - c \\ &= -2(x_i - x_j)^T(u_i - u_j) - \gamma(||x_i - x_j||^2 - D_{\text{safe}}^2) - c \end{aligned} \tag{1}$$

where c as a constant offset is a large positive value to ensure $L_{ij}(x, u)$ is always negative to prevent unintended cancelling-out when being accumulated later. $u_i, u_j \in \mathbb{R}^2$ are the agent's current velocities. γ is the CBF design factor representing how aggressive the pairwise vehicles are [18]. The safety loss function $L_{ij}(x, u)$ [1] represents how close the system is to the boundary of the safe set, or how easily a safety violation could occur, under the assumption that both vehicles will move with piecewise-constant velocity.

[1]Note that this risk evaluation tool does not necessarily require the vehicles to use Control Barrier Function-based controllers. We understand that in the real world vehicles may use different kinds of controllers, yet this does not prevent them from understanding the risk generated from inter-robot interaction via this tool, with the mild but reasonable assumption that information about safety margin and vehicle states is known or observable. Even for vehicles not using CBF-based controllers, it is still possible to learn the parameter γ from observations using machine learning techniques like linear ridge regression [21].

III. METHOD

A. Risk Assessment for Heterogeneous Traffic

To provide the vehicle with a sense of situational awareness of the dynamic environment it is in, we are interested in assessing the accumulated risk a vehicle receives from the environment. Now with $L_{ij}(x, u)$ as a handy tool describing the risk vehicle i faces when interacting with vehicle j, for a scenario involving multiple vehicles, we define the aggregated risk $R_i \in \mathbb{R}$ vehicle i faces posed by all surrounding vehicles. Since we aim to tailor the previously proposed CBF-inspired risk evaluation for mixed-traffic scenarios in autonomous driving as shown in Fig. 1, to explicitly model this difference in social influence from heterogeneous vehicle types, we propose the following social influence weight $w = \begin{bmatrix} w_1 & w_2 & ... & w_N \end{bmatrix} \in \mathbb{R}^N$ for all vehicles in the scene with $w_i = \mathcal{M}(m_i)$, where $m_i \in \mathbb{R}$ is the mass of each vehicle, and $\mathcal{M}$ is a mapping function that maps the vehicle mass to a weight with $\sum_{i=1}^{N} w_i = 1$. The larger the vehicle mass is, the smaller w_i will be. According to the Vehicle Classification Definition of Federal Highway Administration by the U.S. Department of Transportation [25], all vehicles are classified into 13 categories based on the number of axies. For simplicity, we provide the following table of a few common vehicle types for reference, with the data of the approximated weight provided by the Pennsylvania Department of Transportation [26].

TABLE I
VEHICLE WEIGHT APPROXIMATION BY PENNDOT.

Approximated Weight of Traffic Participants (in tons)		
Class I	Motorcycle	0.2
Class II	Sedan	1.5
	SUV	2
Class III or above	Empty Truck	10
	Bus	20
	Heavy Loaded Truck	40

As shown in Table I, Class I refers to motorcycles and Class II consists of small passenger vehicles like Sedan and SUV. For Class III or above, they are mostly commercial vehicles like trucks, buses, and tractors. As an example, if vehicle i is an SUV, and vehicle j is a heavily loaded truck, then $w_i : w_j = 20 : 1$. Next, the accumulated risk R_i the vehicle i receives from the interactive environment with multiple surrounding vehicles is defined as:

$$R_i = \sum_{j=1}^{N} w_j L_{ij}(x, u), \quad \forall j \neq i \tag{2}$$

As the equation suggests, R_i provides a quantitative measure of the amount of risk the vehicle i receives from all surrounding vehicles considering their different vehicle types. Recall that since $L_{ij}(x, u)$ is negative, then the greater mass the vehicle j has, the smaller weight w_j is, therefore the larger value R_i has. The greater R_i is, the more likely a safety violation is to occur. The proposed risk evaluation framework is simple yet effective: 1) R_i grows with the increased number of vehicles in the system, as the environment becomes more complex and challenging; 2) R_i varies depending on the changes of states, including positions and motion of other vehicles as we expected, as it is important to tell how much risk agent i is exposed to even when a collision has not happened yet; 3) With the special mass-related weight design of w, this accumulated risk assessment is augmented by various degrees of social influence of different vehicle types. The underlying idea here is that considering a fully loaded truck and a small passenger vehicle, even if they have the same relative positions and motions compared to the ego vehicle, it is obvious that the heavy truck is considered a higher potential threat. This can be explained by the truck's higher momentum owing to its greater weight, making it considerably more challenging to brake or accelerate compared to the smaller passenger vehicle. This concludes our introduction to the notion of risk that provides the ego vehicle with a means to assess the situation, and we will elaborate on how this understanding can be applied to shape the core design of the socially compliant behavior planning framework.

Given that our objective is to create a socially compliant behavior-planning framework that only minimally alters the existing controller to filter out behaviors incongruent with societal expectations, we must address two pivotal questions: **1) Intervention Condition Definition:** When should our proposed framework intervene and modify the actions of the existing controller? **2) Social Norm Characterization:** How should we define socially compliant behavior that aligns with human expectations? Conversely, how can we identify behaviors that deviate from typical human expectations and are therefore undesirable? These two questions address the "when" and "how" aspects of intervention for the ego robot, forming the cornerstone of our approach toward achieving socially acceptable autonomous vehicle behavior.

B. Reasoning of Accountability

To answer the first question, we argue that intervention is only needed when the existing controller may lead the ego vehicle into a highly risky situation even though a collision has not happened yet. We start by introducing the binary logical operator l:

$$l = \mathbb{I}(\Delta R_e \geq R_{\text{threshold}}) \tag{3}$$

where we use subscript e to denote the ego vehicle. ΔR_e is the difference in the accumulated risk (Eq. 2) the ego vehicle receives from the surrounding environment between two consecutive time steps, and $R_{\text{threshold}}$ is the user-defined threshold value, that defines situations that should engage the ego vehicle's special attention. With l returning true or false, Eq. 3 indicates if there is a significant increase in the accumulated risk the ego vehicle receives from the environment so that it should take a closer look. Such a substantial increase could result from various factors, e.g., a potential sudden acceleration by the following vehicle or unexpected lane changes by vehicles in adjacent lanes.

Once a risky occasion is detected, the ego vehicle should reason about who should be responsible for the surge in

risk during the interaction. To trace accountability, the ego vehicle identifies the one pairwise counterpart j^* that causes the primary surge by examining the change of all pairwise risk assessments between two consecutive time steps:

$$j^* = \arg\max_k \Delta L_{ek}(x,u) \quad \forall k \in \{1,...,N\} \setminus e \tag{4}$$

Then by computing the accountability ϕ for itself and the neighboring vehicle j^*, we trace back to see whose motion $\dot{x}$ is making a higher contribution to the identified risk surge.

$$\phi_e = \frac{\partial L_{ej^*}}{\partial x_e}\dot{x_e} = (-2(u_e - u_{j^*}) - 2\gamma(x_e - x_{j^*}))^T u_e \tag{5}$$

Finally, we determine the vehicle that should be accountable for the risk surge as:

$$k = \arg\max_{k\in\{e,j^*\}} \phi_k = \arg\max_{k\in\{e,j^*\}} \frac{\partial L_{ej^*}}{\partial x_k}\dot{x_k} \tag{6}$$

In this work, we assume that the existing controller $\tilde{u}_e$ in Eq. 7 of the ego vehicle already satisfies the collision-free safety requirement[2], and our goal is to design an interpretable behavior layer to filter out those behaviors that are non-socially compliant. It is important to reason over in what kind of situations the proposed behavior layer should come into play and supersede the existing controllers.

$$\begin{aligned} \tilde{u}_e &= \arg\min_{u_e\in\mathcal{U}} ||u_e - \bar{u}_e||^2 \\ s.t. \quad &||x_e - x_j||^2 \geq D^2_{\text{safe}} \quad \forall j \in \{1,...,N\} \setminus e \end{aligned} \tag{7}$$

where $\bar{u}_e$ is the task-related nominal control command provided by a high-level planner, e.g., a motion planner. Considering that $\tilde{u}_e$ already represents the best response for the ego vehicle to accomplish the designated task while maintaining safety, no intervention will be applied to maximize task performance without unnecessary restrictions. This allows the ego vehicle to execute necessary actions freely. One such example is when the ego vehicle is in the left lane of a highway approaching an exit, and it decides to change lanes in front of a neighboring vehicle already in the right lane. In such a scenario, the ego vehicle is accountable for the significant increase in risk due to this maneuver, but no intervention is required because lane changing is a necessary action. Conversely, if a risk surge is primarily caused by changes in the external environment, such as the behavior of neighboring vehicles, intervention becomes necessary to ensure that the ego vehicle's reactive behavior remains both safe and reasonable.

Therefore, the proposed behavior planning framework is designed only to intervene when $k = j^*$ in Eq. 6, suggesting that it is the pairwise counterpart j^* that causes the risk surge instead of the ego itself. Then the second logical operator is defined as:

$$\delta = \mathbb{I}(\phi_e < \phi_{j^*}) \tag{8}$$

In summary, the essential and sufficient condition for the intervention to happen is: $l \wedge \delta = 1$, namely risk surge is identified and the ego vehicle is not accountable for it.

[2]This is a reasonable assumption considering all the great tools available, including but not limited to Control Barrier Functions [27] and Reachability Analysis [28].

C. Risk-Informed Social Norm Characterization

Now to answer the second question, we draw inspiration from Asimov's "Three Laws of Robotics" [29] which express human expectations governing the behavior of robots. We now use them to guide our design of social norms characterization during the intervention of our framework. For easier representation, we denote the joint control state for each vehicle pair e (ego) and j (neighboring vehicle) without intervention as $\tilde{u}$, consisting of $\tilde{u}_e$ and u_j, and the joint control state after intervention as u^*, consisting u^*_e and u_j.

Asimov's First Law states, "A robot may not injure a human being or, through inaction, allow a human being to come to harm." Since in this work we assume the existing controller $\tilde{u}$ can already guarantee no collision happens, with our notion of risk, we interpret the first law as the expectation of robots' ability to reason over potential risk even when no immediate collision is going to happen: the accumulated risk the ego vehicle **poses to** surrounding vehicles should decrease after intervention u^*, compared to that with the $\tilde{u}$ that the existing controller supplies.

$$\sum_{j\in N\setminus e} w_e L_{ej}(x,u^*) < \sum_{j\in N\setminus e} w_e L_{ej}(x,\tilde{u}) \tag{9}$$

An example involves the ego vehicle being a heavily loaded truck occupying the left lane, with a smaller passenger vehicle following closely. Suddenly, the following vehicle accelerates, significantly reducing the gap between them. In the absence of intervention, the original controller $\tilde{u}$ instructs the ego truck to accelerate in response to maintain safety. However, our intervention u^* directs the ego vehicle to execute a lane change to the right, allowing the following vehicle to pass first. This action is not within the scope of the original controller, which does not consider the increased risk to the human passenger vehicle when closely trailing a heavily loaded truck traveling at a high speed which could put the following vehicle in danger.

The Second Law states, "A robot must obey orders given to it by human beings, except where such orders would conflict with the first law." With the intervention mechanism and the controller designed by a human, the autonomous vehicle adheres to this principle. In light of the recent debate on this second law regarding the responsiveness of robots, Murphy and Woods [30] proposed an alternative second law "A robot must respond to humans as appropriate for their roles", to emphasize that the capability for robots to respond appropriately is more important in human-robot interaction compared to the capability of the autonomy. In the context of shared autonomy in mixed traffic scenarios, the intervention of robot behavior should not be limited solely to situations leading to immediate injury, such as collisions with human-driven cars—precisely the motivation behind this work, which aims to eliminate vehicle behaviors that do not align with social norms when necessary.

The Third Law states, "A robot must protect its own existence as long as such protection does not conflict with the first or second law." Leveraging our notion of risk, we interpret this law as requiring that the accumulated risk the

ego vehicle **receives from** the surrounding vehicles should decrease after intervention u^*, compared to that with $\tilde{u}$.

$$R_e(u^*) < R_e(\tilde{u}) \leftrightarrow \sum_{j \in N \setminus e} w_j L_{ej}(x, u^*) < \sum_{j \in N \setminus e} w_j L_{ej}(x, \tilde{u}) \tag{10}$$

This interpretation forces the ego vehicle to act as a self-preserving entity, prioritizing its own safety. For instance, human drivers do not expect the ego vehicle to execute an abrupt and aggressive lane change in the midst of a closely-following fleet of vehicles, solely to create space for a human driver behind it to pass ahead.

D. Socially-Compliant Control Problem Formulation

Finally, we introduce our proposed Socially-Compliant control problem formulation, which integrates the behavior planning approach with social norm intervention into the original controller design (Eq. 7).

$$\begin{aligned} u_e^* = \arg\min_{u_e \in \mathcal{U}} & ||u_e - \bar{u}_e||^2 \\ & + (l \wedge \delta)(\mu_1 \sum_{j \in N \setminus e} w_e L_{ej}(x, u) + \mu_2 R_e(x, u)) \\ s.t. \quad & ||x_e - x_j||^2 \geq D_{\text{safe}}^2 \quad \forall j \in \{1, ..., N\} \setminus e \\ & l = \mathbb{I}(\Delta R_e \geq R_{\text{threshold}}) \\ & \delta = \mathbb{I}(\phi_e < \phi_{j^*}) \end{aligned} \tag{11}$$

where $\mu_1 > \mu_2 \gg 0$ are two large positive coefficients for the accumulated risk the ego vehicle poses to surrounding vehicles and the accumulated risk it receives from them, prioritizing the first law "no harm to humans". This optimization problem can be solved using a Mixed Integer Quadratic Programming Solver directly or any regular optimization solver by reformulating the problem using the Big-M method [31] for improved computational efficiency.

IV. Simulation & Discussion

We provide three illustrative examples to show the validity and effectiveness of our proposed approach. The existing controller is set to be the same as Eq. 7, namely maintaining a nominated travel speed whenever possible without collisions.

Example 1: We first showcase the performance of our proposed approach in comparison to the existing controller in a two-vehicle scenario, as depicted in Figure 2. We have a scenario plot on the left showing the ego truck on the fast lane with a small passenger vehicle following it. The dashed line represents the decision space of the ego truck in our proposed approach on whether to change its lane and how fast it would like to travel. Distinct ego behavior is observed in three different setups (left, middle, right) on the right.

Next, we demonstrate how heterogeneous road participants can affect the decision of our proposed method in two multi-vehicle scenarios. In both cases, we have one fully loaded truck and two passenger vehicles sharing the same initial states and conditions, with the only difference as the switched vehicle type of the two non-ego vehicles. **Example 2**: In Fig. 3, the small passenger vehicle behind is programmed

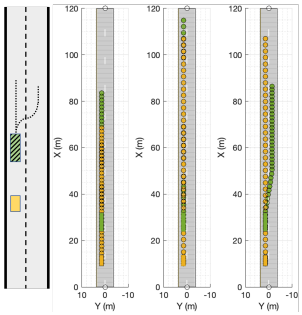

Fig. 2. The scenario illustration and simulation plots for Example 1. The green vehicle is the ego truck and the yellow vehicle is a small passenger vehicle. The three subplots (left, middle, right) on the right introduce different scenario setups with waypoints plotted out for easier understanding. **Left:** The small passenger car maintains a steady speed and doesn't create any pressure or risk for the ego truck. Since no risk surge is detected, our behavior planning framework doesn't need to intervene. Consequently, the ego truck continues to follow its existing controller, and we have $u_i^* = \tilde{u}_i$. **Middle:** The ego truck is solely relying on its existing controller $\tilde{u}_i$ without our behavior planning framework. At a certain point, the small passenger vehicle begins to accelerate, rapidly closing the gap between the two vehicles. Without considering the potential risks to both itself and the human passenger vehicle, $\tilde{u}_i$ is left with no alternative but to instruct the ego truck to also accelerate in order to maintain a safe inter-vehicle distance. **Right:** Here the small passenger vehicle takes the same action as in the middle subplot by accelerating. However, the difference is that the ego truck is equipped with our proposed algorithm. As the gap between the two vehicles rapidly shrinks, the increased risk to the ego truck triggers the accountability trace, leading the ego truck to hold the small passenger vehicle accountable. Subsequently, the intervention mechanism is activated. In evaluating various choices available to the ego truck and considering their alignment with typical human expectations, it decides to temporarily deviate from its nominal controller. It does so by executing a lane change to the right lane, allowing the small passenger vehicle to pass first. This decision is a wiser one compared to the scenario in the middle subplot, as the ego truck chooses not to jeopardize the safety of both itself and the small passenger vehicle. At the same time, it strives to adhere to the task-related nominal controller to the greatest extent possible.

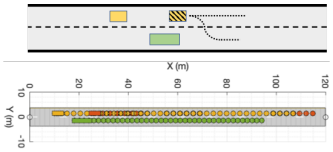

Fig. 3. The ego vehicle is the yellow passenger vehicle with black strips in the scenario illustration, corresponding to the red vehicle in the simulation illustration for clearer visualization. There is one small passenger vehicle following the ego vehicle, and a large truck in the adjacent lane.

to aggressively accelerate in a very short time frame. Once again, our proposed framework activates the intervention mechanism, prompting the ego passenger vehicle to consider its available options. In contrast to the previous example, it recognizes that executing a lane change is not the most favorable solution here. This is because changing lanes would introduce a higher accumulated risk the ego vehicle poses to surrounding vehicles, especially to the truck in the adjacent lane, surpassing the accumulated risk it generates when staying in the current lane and accelerating to maintain a safe distance from the following passenger vehicle. **Example 3**: However, the situation takes a different turn when the vehicle types of the two non-ego vehicles are swapped as shown in Fig. 4. Now, with a fully loaded truck following the

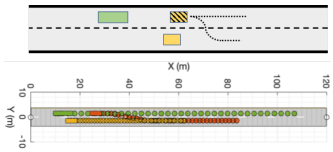

Fig. 4. Example 3: There is one large truck following the ego vehicle and a small passenger vehicle in the adjacent lane. This recommendation is based on the fact that the cumulative risk the ego vehicle incurs, amplified by the weight-related mass, greatly exceeds the risk associated with changing lanes in front of the small passenger vehicle in the adjacent lane. This decision, even if it involves a temporary deviation from its existing controller, is deemed more prudent.

ego vehicle and exhibiting aggressive acceleration, despite all three vehicles sharing identical state configurations, including positions and motions, our proposed method advises the ego passenger vehicle to execute a lane change.

Conclusion We extend our CBF-based accumulated risk evaluation to realistic highway driving scenarios that consider the heterogeneity in vehicle types to embed their potentially different levels of social influence into the notion of risk. We achieve this by incorporating accountability tracing and social norm characterization into mathematical expressions linked to the concept of risk. This framework is designed to facilitate autonomous vehicles in exhibiting behaviors that align with common human intuitions, all while guaranteeing a minimum level of performance compared to their existing controllers. Our approach draws inspiration from Isaac Asimov's "Three Laws of Robotics," offering fundamental interpretations of our proposed notion of risk that can be broadly applied to other robotics applications, promoting socially compliant robot behavior generation.

REFERENCES

[1] L. Claussmann, M. Revilloud, D. Gruyer, and S. Glaser, "A review of motion planning for highway autonomous driving," *Transactions on Intelligent Transportation Systems*, vol. 21, pp. 1826–1848, 2019.

[2] J. Chen, W. Zhan, and M. Tomizuka, "Autonomous driving motion planning with constrained iterative lqr," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 2, pp. 244–254, 2019.

[3] Y. Lyu, W. Luo, and J. M. Dolan, "Probabilistic safety-assured adaptive merging control for autonomous vehicles," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 10 764–10 770.

[4] M. Althoff, D. Althoff, D. Wollherr, and M. Buss, "Safety verification of autonomous vehicles for coordinated evasive maneuvers," in *2010 IEEE Intelligent Vehicles Symposium*. IEEE, 2010, pp. 1078–1083.

[5] J. K. Choi and Y. G. Ji, "Investigating the importance of trust on adopting an autonomous vehicle," *International Journal of Human-Computer Interaction*, vol. 31, no. 10, pp. 692–702, 2015.

[6] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of experimental social psychology*, vol. 52, pp. 113–117, 2014.

[7] L. Oliveira, K. Proctor, C. G. Burns, and S. Birrell, "Driving style: How should an automated vehicle behave?" *Information*, vol. 10, no. 6, p. 219, 2019.

[8] F. Riaz, S. Jabbar, M. Sajid, M. Ahmad, K. Naseer, and N. Ali, "A collision avoidance scheme for autonomous vehicles inspired by human social norms," *Computers & Electrical Engineering*, vol. 69, pp. 690–704, 2018.

[9] P. Hang, C. Lv, Y. Xing, C. Huang, and Z. Hu, "Human-like decision making for autonomous driving: A noncooperative game theoretic approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 4, pp. 2076–2087, 2020.

[10] Z. Huang, H. Liu, J. Wu, and C. Lv, "Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving," *Transactions on neural networks and learning systems*, 2023.

[11] J. Grover, Y. Lyu, W. Luo, C. Liu, J. Dolan, and K. Sycara, "Semantically-aware pedestrian intent prediction with barrier functions and mixed-integer quadratic programming," *IFAC-PapersOnLine*, vol. 55, no. 41, pp. 167–174, 2022.

[12] J. Ding, L. Li, H. Peng, and Y. Zhang, "A rule-based cooperative merging strategy for connected and automated vehicles," *Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3436–3446, 2019.

[13] A. Aksjonov and V. Kyrki, "Rule-based decision-making system for autonomous vehicles at intersections with mixed traffic environment," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 660–666.

[14] L. Zhou and P. Tokekar, "Risk-aware submodular optimization for multirobot coordination," *IEEE Transactions on Robotics*, 2022.

[15] E. Scukins and P. Ögren, "Using reinforcement learning to create control barrier functions for explicit risk mitigation in adversarial environments," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 10 734–10 740.

[16] Y. Lyu, W. Luo, and J. M. Dolan, "Risk-aware safe control for decentralized multi-agent systems via dynamic responsibility allocation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1–8.

[17] Y. Lyu, J. M. Dolan, and W. Luo, "Decentralized safe navigation for multi-agent systems via risk-aware weighted buffered voronoi cells," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, ser. AAMAS '23. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2023, p. 1476–1484.

[18] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *2019 18th European control conference (ECC)*. IEEE, 2019, pp. 3420–3431.

[19] J. Zeng, B. Zhang, and K. Sreenath, "Safety-critical model predictive control with discrete-time control barrier function," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 3882–3889.

[20] S. He, J. Zeng, B. Zhang, and K. Sreenath, "Rule-based safety-critical control design using control barrier functions with application to autonomous lane change," 2021.

[21] Y. Lyu, W. Luo, and J. M. Dolan, "Adaptive safe merging control for heterogeneous autonomous vehicles using parametric control barrier functions," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 542–547.

[22] ——, "Responsibility-associated multi-agent collision avoidance with social preferences," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2022, pp. 3645–3651.

[23] W. Luo, W. Sun, and A. Kapoor, "Multi-robot collision avoidance under uncertainty with probabilistic safety barrier certificates," *Advances in Neural Information Processing Systems*, vol. 33, pp. 372–383, 2020.

[24] S. Van Koevering, Y. Lyu, W. Luo, and J. Dolan, "Provable probabilistic safety and feasibility-assured control for autonomous vehicles using exponential control barrier functions," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 952–957.

[25] F. H. Administration, "Verification, refinement, and applicability of long-term pavement performance vehicle classification rules."

[26] D. of Transportation of Pennsylvania, "Approxinate vehicle weights."

[27] A. D. Ames, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs with application to adaptive cruise control," in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 6271–6278.

[28] A. B. Kurzhanski and P. Varaiya, "Ellipsoidal techniques for reachability analysis," in *International workshop on hybrid systems: Computation and control*. Springer, 2000, pp. 202–214.

[29] I. Asimov, "Three laws of robotics."

[30] R. Murphy and D. D. Woods, "Beyond asimov: The three laws of responsible robotics," *IEEE intelligent systems*, vol. 24, no. 4, pp. 14–20, 2009.

[31] I. Griva, S. G. Nash, and A. Sofer, *Linear and Nonlinear Optimization 2nd Edition*. SIAM, 2008.