2024 International Conference on Assured Autonomy (ICAA) | 979-8-3315-2101-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICAA64256.2024.00012

Who is Responsible? Explaining Safety Violations in Multi-Agent Cyber-Physical Systems

Luyao Niu[1], Hongchao Zhang[2], Dinuka Sahabandu[1],
Bhaskar Ramasubramanian[3], Andrew Clark[2], Radha Poovendran[1]

***Abstract*—Multi-agent cyber-physical systems are present in a variety of applications. Agent decision-making can be affected due to errors induced by uncertain, dynamic operating environments or due to incorrect actions taken by an agent. When an erroneous decision that leads to a violation of safety is identified, assigning responsibility to individual agents is a key step towards preventing future accidents. Current approaches to carrying out such investigations require human labor or high degree of familiarity with operating environments. Automated strategies to assign responsibility can achieve significant reduction in human effort and associated cognitive burden.**

In this paper, we develop an automated procedure to assign responsibility for safety violations to actions of any single agent in a principled manner. We ground our approach on reasoning about safety violations in road safety. When provided with an instance of a safety violation, we use counterfactual reasoning to create alternate scenarios that determine how different outcomes might have been achieved if a specific action or set of actions was replaced by another action or set of actions. We devise a metric called the *degree of responsibility (DoR)* for each agent. The DoR uses the Shapley value to quantify the relative contribution of each agent to the observed safety violation, thus serving as a basis to explain and justify future decisions. We devise both heuristic techniques and methods based on the structure of agent interactions to improve scalability of our solution as the number of agents increases. We consider three instances of safety violations from the National Highway Traffic Safety Administration (NHTSA). We carry out experiments using representations of the three scenarios using the CARLA urban driving simulator. Our results indicate that the DoR enhances explainability of decision-making and assigning accountability for actions of agents and their consequences.

I. Introduction

The operation of complex cyber and cyber-physical systems (CPS) such as autonomous cars relies heavily on the seamless integration of physical components with software-based decision making algorithms and procedures. When CPS are deployed in the real-world, over the horizon of its operation, each system will have to solve a sequential decision making problem in order to 'optimally' satisfy desired objectives (e.g., safety, reachability) in uncertain environments. The safe operation of CPS is enforced as a constraint, typically on the state of the system [1]–[6]. Examples of safety constraints include (i) any pair of agents should be at least some distance d apart in order to avoid collisions, or (ii) no agent should enter a specified 'forbidden' region in the environment.

There is a large body of work that has developed solutions with guarantees of safety during the design or operation phases of multi-agent CPS [1]–[9]. Two well-established directions of research in this domain are verification during design [7], [8], [10] and synthesizing safety-critical controllers during deployment [1]–[6]. However, uncertainties or changes in the operating environment or erroneous agent behavior might result in safety violations. For example, a controller synthesized to be safe for a car driven in clear weather may not be safe when it is raining [11], [12].

Devising strategies to mitigate safety violations will be especially crucial for successful large-scale deployment of multi-agent CPS. A particular multi-agent CPS that we consider is a road network with emphasis on urban driving. The reason for this is twofold: (i) erroneous agent decisions (i.e., incorrect actions by vehicle drivers) have been known to result in safety violations (i.e., cause an accident), and (ii) road traffic accidents have been known to be a leading cause of death and billions of dollars of economic losses in the US [13], [14].

When errors in decision making by agents lead to safety violations, a key challenge is to identify and assign responsibility to actions of individual agents. This will inform development of measures to prevent future accidents, e.g., by correcting flaws in perception/control algorithms, or introducing new safety rules. However, current approaches to carrying out such investigations are either application/instance-specific [15] or require human labor [16], [17], which can incur large costs. In some cases, a high level of familiarity with the operating environment [18] might also be required, which can place significant cognitive burden on human analysts. Designing automated procedures to assign responsibility will aid in reducing the cognitive overload on human operators. Further, such procedures can serve as an auditable basis to explain and justify future decisions, including those that might be made by human analysts.

In this paper, we develop foundations of an automated procedure to assign responsibility to actions of any single agent relative to other agents after a safety violation in a principled manner. When presented with an example (in our setup, a trajectory) of a safety violation, we use *counterfactual*

[1]Department of Electrical and Computer Engineering, University of Washington, Seattle, WA, USA. {luyaoniu, sdinuka, rp3}@uw.edu

[2]Department of Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, MO, USA. {hongchao, andrewclark}@wustl.edu

[3]Electrical and Computer Engineering, Western Washington University, Bellingham, WA, USA. ramasub@wwu.edu

This work was supported by the Office of Naval Research grant N0014-23-1-2386, National Science Foundation grants CNS 1941670, CNS 2153136, and Air Force Office of Scientific Research grants FA9550-22-1-0054 and FA9550-23-1-0208.

979-8-3315-2101-1/24/$31.00 ©2024 IEEE
DOI 10.1109/ICAA64256.2024.00012

reasoning [19], [20] to identify and explain agent roles and responsibilities. Counterfactual reasoning provides a way of creating alternate scenarios, i.e., *counter to the facts*, that seek to determine how different an outcome might have been if a specific action or set of actions was replaced by another action or set of actions. By systematically comparing outcomes in the *counterfactual worlds* with actual observed outcomes, we will be able to identify which agents were responsible for the observed safety violation, relative to other agents.

We reason about responsibility of individual agents through a multi-agent Markov decision process (MMDP), a trajectory of which has been observed to violate safety. The primary challenge in this setting is that we only have access to one trajectory that violated the safety constraint, and we lack information on how agents would behave when taking some alternative actions. We propose to mitigate this challenge by formulating a coalitional game on the MMDP, where a coalition consists of a subset of agents who choose their actions in a counterfactual world according to a safe control policy, i.e., a policy that minimizes probability of violating the safety constraint. For a given coalition, we construct a utility function that captures whether agents within the coalition had alternate actions available to them that would have prevented violation of safety. We devise a metric that we term the *degree of responsibility (DoR)*. The DoR leverages insights from the Shapley value [21] from game theory to quantify the relative contribution of each agent to the observed safety violation. Different from causal reasoning [22], our DoR is not necessarily a binary value when assigning responsibilities.

To scale our solution to large numbers of interacting agents, we develop both heuristic solutions and methods based on the structure of agent interactions to improve efficiency of DoR computation. We design an algorithm to characterize the marginal contribution of an agent to the safety violation, and use this algorithm to isolate a subset of potentially responsible agents. We also show that when the MMDP satisfies an exponential decay property [23], the DoR of each agent can be obtained with lower computational and memory complexities.

We ground our approach on providing explanations for safety violations by automobiles, motivated by the economic and social impacts of road traffic accidents [13], [14]. We consider three instances of safety violations from the National Highway Traffic Safety Administration (NHTSA) [24]. We carry out extensive experiments using representations of the three scenarios within the CARLA urban driving simulator [25] to identify and quantify agent responsibility for a safety violation using counterfactual reasoning. Our results indicate that our DoR metric enhances explainability of decision-making and assigning accountability for actions of agents and their consequences. Moreover, comparing the DoRs of agents yields an ordered list of responsibilities, which can serve as a sound basis to explain and justify subsequent decisions, including those potentially taken by humans. In these scenarios, the actions taken by agents are coupled and their epistemic states are often unknown, making existing reasoning techniques [26] inadequate.

The remainder of this paper is organized as follows: Sec. II discusses related work. We formulate the problem studied in Sec. III and detail our solution approach in Sec. IV. Sec. V presents a scalable way to compute the DoR when there are a large number of agents. We perform experiments to validate our approach in Sec. VI and conclude the paper in Sec. VII.

II. Related Work

Guaranteeing safety in multi-agent systems (MAS) has been extensively studied. Typical solutions include control-theoretic based approaches and learning-based approaches. When the agents follow known dynamics or incur bounded uncertainty in the worst-case, control-theoretic approaches can be applied to synthesize controllers for the agents [1]–[6]. When the dynamics followed by the agents are unknown, learning-based approaches [8], [27]–[34] can be utilized to learn the controllers with safety guarantees. The aforementioned studies mainly focus on designing controllers with safety guarantees for MAS or verifying safety. In this paper, we consider the scenarios where safety is observed to be violated by the multi-agent system. Our goal is to quantify the individual agent's responsibility for the observed safety violation.

Although there have been extensive studies on guaranteeing safety of multi-agent CPS, safety may still be violated in practical applications due to uncertainties, faults, and errors. There are two major categories of approaches to investigate safety violations. The first category utilizes data records on safety violations to identify key risk factors. For example, some factors such as weather, lighting, and roadway surfaces are identified in [35]. In [36], statistical analysis is performed to analyze autonomous vehicle collisions, with particular emphasis on collision types and maneuvers as well as errors of drivers of conventional vehicles. Crash narratives are utilized in [37] to identify the factors contributing to autonomous vehicle crashes. The second category focuses on assisting the accident investigators to explain the safety violations. In [16], a why-because list is constructed by the investigation team to explain safety violations of social robots. In [17], the authors developed a tree-based representation to explain the behaviors of autonomous vehicles. A data pipeline is proposed in [18] for investigators of autonomous vehicle crashes to reconstruct the scene of accidents using raw data from sensors mounted by the vehicles. In contrast, this paper aims to develop an automated procedure to quantify the responsibilities of agents in multi-agent CPS for safety violations, which has not been investigated in the aforementioned studies.

In [15], responsibilities for crossroad vehicle collisions are assessed based on traffic rules. In this paper, we leverage the Shapley value [21] to develop an explainable metric to quantify the responsibility of each individual agent for the safety violation. The Shapley value has also been adopted in artificial intelligence to explain why machine models produce certain outputs for given inputs [38]. However, quantifying responsibility in multi-agent CPS studied in this paper presents unique challenges, distinct from [15], [38]. That is, quantification of each agent's responsibility in our work requires

decoupling and isolating not only effects of actions taken by other agents but also the influence of its own prior actions.

III. System Model and Problem Formulation

We consider a multi-agent CPS consisting of a collection of agents, denoted $\mathcal{N} = \{1, \ldots, N\}$. The dynamics and interactions among the agents are modeled as a multi-agent Markov decision process (MMDP), which is defined as follows.

Definition 1 (Multi-Agent Markov Decision Process (MMDP))**.** *A multi-agent Markov decision process $\mathbb{M}$ is a tuple $\mathbb{M} = (\{\mathcal{S}_i\}_{i=1}^N, \{\mathcal{A}_i\}_{i=1}^N, Pr)$, where $\mathcal{S}_i$ is a finite set of states of agent $i \in \mathcal{N}$, $\mathcal{A}_i$ is a finite set of actions available to agent i, and $Pr : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition function where $\mathcal{S} = \prod_{i=1}^N \mathcal{S}_i$ and $\mathcal{A} = \prod_{i=1}^N \mathcal{A}_i$ are the joint state and action spaces of all agents, and $\Delta(\mathcal{S})$ denotes the set of probability distributions over set $\mathcal{S}$. The probability of transitioning from state $s = (s_1, \ldots, s_N)$ to state $s' = (s'_1, \ldots, s'_N)$ when the agents take actions $a = (a_1, \ldots, a_N)$ is written as $Pr(s'|s, a)$.*

We define the *joint non-stationary deterministic policy* of the agents on the MMDP $\mathbb{M}$ as $\pi : \mathcal{S} \times \mathbb{Z} \to \mathcal{A}$, which specifies a joint action $a \in \mathcal{A}$ of all agents for any joint state $s \in \mathcal{S}$ and stage $t \in \mathbb{Z}$. A *non-stationary deterministic policy* of an agent i is a mapping $\pi_i : \mathcal{S} \times \mathbb{Z} \to \mathcal{A}_i$ such that $\pi_i(s, t) \in \mathcal{A}_i$ for any joint state $s \in \mathcal{S}$ and stage $t \in \mathbb{Z}$. Thus, for any joint state $s \in \mathcal{S}$, we can represent the joint policy $\pi(s, t)$ as a tuple $\pi(s, t) = (\pi_1(s, t), \ldots, \pi_N(s, t))$.

Given the MMDP $\mathbb{M}$, a *finite path* ρ on $\mathbb{M}$ is a sequence of state-action pairs of finite length. Given a finite path $\rho = s^0, a^0, s^1, \ldots, a^{T-2}, s^{T-1}$, the path of each individual agent $i \in \mathcal{N}$ can be recovered as $\rho_i = s_i^0, a_i^0, s_i^1, \ldots, a_i^{T-2}, s_i^{T-1}$. In the remainder of this paper, we denote the state s^t (resp. s_i^t) visited by path ρ (resp. ρ_i) as ρ^t (resp. ρ_i^t). Given a time horizon T, a joint policy π, and initial state s^0 of the MMDP $\mathbb{M}$, it induces a collection of paths, where each path $\rho = s^0, a^0, s^1, \ldots, a^{T-2}, s^{T-1}$ satisfies $a^t = \pi(s^t, t)$ and $Pr(s^{t+1}|s^t, a^t) > 0$ for all $t = 0, \ldots, T-2$.

We illustrate the MMDP in Definition 1 using an example.

Example 1. *We consider a road segment consisting of two lanes modeled by a discrete set of locations $\mathcal{L} = \{l_1, \ldots, l_8\}$, as shown in Fig. 1. A set of autonomous vehicles $\mathcal{N} = \{1, 2, 3, 4\}$ shares the road segment. Each state $s_i \in \mathcal{S}_i$ of a vehicle i represents its location, where $\mathcal{S}_i = \mathcal{L}$ for all $i \in \mathcal{N}$. Each vehicle i could take certain actions from a set $\mathcal{A}_i$ to transit among the locations. Due to the heterogeneous makes, types, and models, the vehicles follow different transition probabilities $Pr_i : \mathcal{S}_i \times \mathcal{A}_i \to \Delta(\mathcal{S}_i)$, where $Pr(s'_i|s_i, a_i)$ represents the probability of transitioning from a location s_i to its neighboring location s'_i when the vehicle i takes action a_i. Since the vehicles are transition independent, the joint transition probability of all agents is expressed as $Pr(s'|s, a) = \prod_{i \in \mathcal{N}} Pr_i(s'_i|s_i, a_i)$, where $s' = (s_1, \ldots, s_N)$, $s = (s_1, \ldots, s_N)$, and $a = (a_1, \ldots, a_N)$. A path ρ in this example then represents a sequence of joint locations and actions of all vehicles. A policy π_i of vehicle i specifies the action that should be taken by the vehicle given the current stage t and joint location s of all vehicles.*

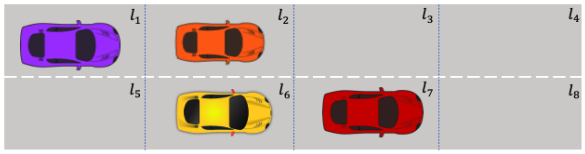


Fig. 1: This figure presents a road segment consisting of two lanes. The road segment is discretized into eight discrete locations. Four vehicles share the road segment.

We assume that there is a subset of unsafe states $\hat{\mathcal{S}} \subset \mathcal{S}$ that should be avoided by the agents. In addition, we assume that $Pr(s'|s, a) = 0$ for all $s' \notin \hat{\mathcal{S}}, s \in \hat{\mathcal{S}}$, and $a \in \mathcal{A}$. This assumption indicates that once an agent reaches an unsafe state, the safety constraint is violated and cannot be recovered.

In this paper, we focus on the case where an instance of safety violation is observed. Without loss of generality, we focus on the case where a finite path ρ of length T is observed such that $\rho^T \in \hat{\mathcal{S}}$. We investigate the following problem.

Problem 1. *Consider a collection of agents modeled by an MMDP where a subset of states $\hat{\mathcal{S}} \subset \mathcal{S}$ is unsafe. Suppose a path ρ that reached $\hat{\mathcal{S}}$ is observed. Characterize the subset of agents that are responsible for the safety violation. Furthermore, if an agent is considered responsible, quantify to what extent this agent is responsible for the safety violation.*

We illustrate Problem 1 using Example 1.

Example 1 (Continued)**.** *Suppose the autonomous vehicles shown in Fig. 1 are required to avoid colliding with each other. Then, the set of unsafe states $\hat{\mathcal{S}}$ can be expressed as $\hat{\mathcal{S}} = \{(s_1, \ldots, s_N) : \exists i, j \in \mathcal{N} \text{ s.t. } s_i = s_j\}$. Suppose a path ρ is observed, such that there exists a stage $t \in \{0, \ldots, T-1\}$ in which vehicles 2 (orange) and 3 (yellow) collide, i.e., $s_2 = s_3$ holds at some stage t. Once these vehicles collide, they will stop there and the safety constraint cannot be recovered. Then the problem of interest is to quantify the vehicles' responsibilities for the collision.*

IV. Solution Approach

This section details development of a concept that we term *Degree of Responsibility (DoR)* to measure individual agents' responsibility for the safety violation observed in path ρ.

A. Overview

Our goal is to quantify the responsibility of each individual agent for the safety violation observed in path ρ. We determine responsibility of an agent by examining whether it could have taken an action or actions different than those observed in path ρ in order to prevent the safety violation. Since we only have access to the observed path ρ, we lack knowledge of how the agents would behave when they take different actions. We

address this lack of information by constructing a collection of *counterfactual worlds*. Each counterfactual world represents a different hypothetical scenario.

In each counterfactual world, a subset of agents, denoted $\mathcal{Y} \subset \mathcal{N}$, chooses actions from a safe policy, i.e., a policy that minimizes the probability of violating the safety constraint. We formalize such a counterfactual world by formulating it as a coalitional game [39], where a subset of agents forms a coalition $\mathcal{Y}$ and follow the safe policy at a particular stage t, while the remaining agents in $\mathcal{N} \setminus \mathcal{Y}$ follow the same actions as they have done in the t-th stage of path ρ. We characterize each counterfactual world by constructing a utility function representing the probability of reaching the unsafe states.

Given the utility functions of the counterfactual worlds, quantifying agents' responsibilities then reduces to a problem of distributing the utilities among the agents. One solution concept to distribute the utility is via the Shapley value [21]. The Shapley value has been adopted in multiple engineering applications [38], [40] to evaluate the contribution of individual agents in a coalition. This motivates us to quantify agent responsibility by splitting utility functions in all counterfactual worlds to agents using the Shapley value.

B. Degree of Responsibility

In this subsection, we discuss how to construct the collection of counterfactual worlds. We then propose the concept of DoR to distribute utilities of these counterfactual worlds among agents to measure their responsibilities for safety violation.

We denote the collection of counterfactual worlds as $\{\mathcal{C}^t_{\mathcal{Y}} : \mathcal{Y} \subseteq \mathcal{N}, t \in \{0, \dots, T-1\}\}$. In each counterfactual world $\mathcal{C}^t_{\mathcal{Y}}$, a subset of agents forms a coalition $\mathcal{Y} \subseteq \mathcal{N}$ at stage $t \in \{0, \dots, T-1\}$ to collaboratively guarantee satisfaction of the safety constraint, i.e., avoid reaching unsafe states $\hat{\mathcal{S}}$. Each counterfactual world $\mathcal{C}^t_{\mathcal{Y}}$ is constructed as follows:

- The joint state of agents in $\mathcal{C}^t_{\mathcal{Y}}$, denoted $\tilde{s}^t$, is identical to the t-th stage of the observed path ρ, i.e., $\tilde{s}^t = \rho^t$.
- A subset of agents forms a coalition $\mathcal{Y}$ and takes a certain action $\tilde{a}^t_i = \pi_i(\tilde{s}^t, t)$ at stage t according to a non-stationary policy π_i for all $i \in \mathcal{Y}$. The policies $\{\pi_i : i \in \mathcal{Y}\}$ minimizes the probability of reaching the unsafe states starting from stage t.
- Each agent $j \notin \mathcal{Y}$ follows the same action that it has taken along ρ at stage t, i.e., $\tilde{a}^t_j = a^t_j$ for all $j \notin \mathcal{Y}$.
- All agents follow a joint non-stationary policy $\tilde{\pi}$ for all stages $t' \in \{t+1, \dots, T-1\}$ to minimize the probability of reaching the unsafe states.

We illustrate the construction of counterfactual worlds by continuing our discussion from Example 1.

Example 1 (Continued)**.** *We construct the counterfactual world $\mathcal{C}^0_{\{1\}}$ as follows. In this counterfactual world, the joint state of all agents is $\tilde{s}^0 = \rho^0$. In addition, we have $\mathcal{Y} = \{1\}$. Thus, vehicle 1 will follow policy π_1 at stage 0 that might be different from its action taken along the path ρ. All other vehicles in $\mathcal{N} \setminus \{1\} = \{2, 3, 4\}$ will follow actions that they have taken at stage 0 in path ρ. For stages $t' \in \{1, \dots, T-1\}$, all agents will follow a joint non-stationary policy that minimizes the probability of reaching the unsafe states.*

There are two major differences between each counterfactual world $\mathcal{C}^t_{\mathcal{Y}}$ and the observed path ρ. First, each agent i in coalition $\mathcal{Y}$ will take actions following a different policy π_i. This allows us to characterize whether safety can be preserved when agents in $\mathcal{Y}$ could have taken some alternative actions. Moreover, the agents that are not in coalition $\mathcal{Y}$ are allowed to take actions observed at the stage t in the counterfactual world $\mathcal{C}^t_{\mathcal{Y}}$. For all stages $t' = t+1, \dots, T-1$, all agents will follow a policy $\tilde{\pi}$ that minimizes the probability of reaching the unsafe states. Thus, we can isolate the effect of actions taken by agents in $\mathcal{N} \setminus \mathcal{Y}$ at stage t of the observed path ρ.

Leveraging the differences between the counterfactual worlds and the observed path, we develop a utility function for each counterfactual world to characterize the probability of reaching the unsafe states as follows:

$$r(C^t_{\mathcal{Y}}; \{\pi_i\}_{i\in\mathcal{Y}}, \tilde{\pi}) = \min_{\substack{\pi_i : i\in\mathcal{Y} \\ \tilde{\pi}}} P(\{t, \dots, T-1\} \rightsquigarrow \hat{\mathcal{S}} | \mathcal{C}^t_{\mathcal{Y}}), \quad (1)$$

where $\{t, \dots, T-1\} \rightsquigarrow \hat{\mathcal{S}}$ denotes the event that there exists some stage $t' \in \{t, \dots, T-1\}$ such that the joint state $\tilde{s}^{t'} \in \hat{\mathcal{S}}$ in the counterfactual world $\mathcal{C}^t_{\mathcal{Y}}$, and $P(\cdot|\mathcal{C}^t_{\mathcal{Y}})$ represents the probability evaluated in the counterfactual world $\mathcal{C}^t_{\mathcal{Y}}$. In Eq. (1), the probability of reaching unsafe states is minimized over two policies: (i) the policies π_i of each agent i in coalition $\mathcal{Y}$ at stage t, and (ii) the policies $\tilde{\pi}$ of all agents $i \in \mathcal{N}$ at stages $t' \in \{t+1, \dots, T-1\}$. Therefore, the utility function in Eq. (1) characterizes the probability of reaching unsafe states if agents in a coalition $\mathcal{Y}$ could have taken some alternative actions. Consequently, this allows us to quantify to what extent actions taken by agents outside the coalition, i.e., in $\mathcal{N} \setminus \mathcal{Y}$ at stage t of path ρ contributes to the observed safety violation.

We characterize the utility function in Eq. (1) as follows.

Proposition 1. *Let $\mathcal{C}^t_{\mathcal{Y}}$ and $\mathcal{C}^t_{\mathcal{Y}'}$ be two counterfactual worlds such that $\mathcal{Y} \subseteq \mathcal{Y}'$. Then*

$$r(C^t_{\mathcal{Y}}; \{\pi_i\}_{i\in\mathcal{Y}}, \tilde{\pi}) \geq r(C^t_{\mathcal{Y}'}; \{\pi'_i\}_{i\in\mathcal{Y}}, \tilde{\pi}'), \quad (2)$$

where $\{\pi_i\}_{i\in\mathcal{Y}}$ and $\tilde{\pi}$ are associated with $\mathcal{C}^t_{\mathcal{Y}}$, and $\{\pi'_i\}_{i\in\mathcal{Y}}$ and $\tilde{\pi}'$ are associated with $\mathcal{C}^t_{\mathcal{Y}'}$.

We note that the counterfactual worlds are independent of each other. Therefore, given the utility function in Eq. (1), we can characterize the probability of reaching the unsafe states $\hat{\mathcal{S}}$ if some agents in $\mathcal{Y}$ could have taken some alternative actions at all stages as follows:

$$u(\mathcal{Y}) = \sum_{t=0}^{T-1} r(C^t_{\mathcal{Y}}; \{\pi_i\}_{i\in\mathcal{Y}}, \tilde{\pi}), \quad (3)$$

Eq. (3) can then be utilized to quantify contribution of actions taken by agents in $\mathcal{N} \setminus \mathcal{Y}$ at all stages $t = 0, \dots, T-1$ in path ρ towards the observed safety violation.

Given Eq. (3), we use the Shapley value [21] to distribute the total amount of utility $u(\mathcal{Y})$ among agents as follows:

$$\phi_i = \frac{1}{|\mathcal{N}|} \sum_{\mathcal{Y} \subseteq \mathcal{N} \setminus \{i\}} \binom{\mathcal{N}}{|\mathcal{N}| - |\mathcal{Y}| - 1}^{-1} (u(\mathcal{Y} \cup \{i\}) - u(\mathcal{Y})). \tag{4}$$

Eq. (4) considers all possible orders in which the agents can join a coalition, and distributes the utility to agent i as its marginal contribution, i.e., $u(\mathcal{Y} \cup \{i\}) - u(\mathcal{Y})$, averaged over all possible orders.

We finally normalize ϕ_i among all the agents, and define the normalized value as the ***degree of responsibility (DoR)*** of agent i, which is given as

$$\psi_i = \frac{\phi_i}{\sum_{i \in \mathcal{N}} \phi_i}. \tag{5}$$

The DoR can then be used to quantify the responsibility of agent i for violating the safety constraint. By Proposition 1, we have that ϕ_i in Eq. (4) is non-positive for all agent $i \in \mathcal{N}$. Hence the DoRs satisfy that $\psi_i \in [0, 1]$ for all agent $i \in \mathcal{N}$ and $\sum_{i \in \mathcal{N}} \psi_i = 1$. This allows us to use the DoR to compare and rank the responsibilities of different agents for the safety violation. We illustrate the DoR using Example 1.

Example 1 (Continued)**.** *We consider the set of counterfactual worlds* $\{\mathcal{C}^t_{\{1\}}\}^{T-1}_{t=0}$*. When* $\mathcal{Y} = \{1\}$*, Eq.* (3) *models the probability of reaching unsafe states in these counterfactual worlds. Note that the vehicle* 1 *follows the policy* $\pi_1(\rho^t, t)$ *at stage* t *in each counterfactual world* $\mathcal{C}^t_{\{1\}}$*. Therefore, Eq.* (3) *characterizes the influence of actions taken by vehicle* 1 *at different stages* $t = 0, \ldots, T-1$*. In Eq.* (4)*, the term* $u(\mathcal{Y} \cup \{2\}) - u(\mathcal{Y})$ *quantifies the effect of actions taken by vehicle* 2 *at different stages, given that vehicle* 1 *will follow policy* π_1 *at each stage. By taking all possible permutations* $\mathcal{Y} \subseteq \mathcal{N} \setminus \{2\}$ *into consideration and normalizing among all agents, Eq.* (5) *yields the DoR of vehicle* 2 *to be* 0.5*.*

C. Computation of Policies in the Counterfactual World

To compute the DoRs for all agents, Eq. (1)-(5) require to (i) construct a policy π_i for each agent $i \in \mathcal{Y}$ at stage t that minimizes the probability of reaching unsafe states $\hat{\mathcal{S}}$, (ii) construct a policy $\tilde{\pi}$ for all agents at stages $t' = t+1, \ldots, T-1$, and (ii) evaluate the probability of reaching the unsafe states $\hat{\mathcal{S}}$ in each counterfactual world. We refer to the policies π_i followed by the agents in $\mathcal{Y}$ and $\tilde{\pi}$ in the counterfactual worlds as the *safe policies*. In what follows, we discuss how to compute the safe policies and how to evaluate the probability of reaching the unsafe states in each counterfactual world.

To compute safe policies and evaluate the probability of reaching the unsafe states $\hat{\mathcal{S}}$ in each counterfactual world, we define a Q-function for any given joint state-action pair as:

$$Q(s, a, 0) = \begin{cases} 0, & \forall a \text{ if } s \notin \hat{\mathcal{S}} \\ 1, & \forall a \text{ if } s \in \hat{\mathcal{S}} \end{cases}, \tag{6}$$

$$Q(s, a, t) = \sum_{s' \in \mathcal{S}} Pr(s'|s, a) \min_{a'} Q(s', a', t-1), \ \forall t > 1. \tag{7}$$

The Q-function specifies the probability of reaching the unsafe states in t stages when the agents start from state s and take joint action a. By the definition of the Q-function, we have that the policy π_i for each agent $i \in \mathcal{Y}$ can be computed by minimizing $Q(\rho^t, \tilde{a}^t, T-t)$ over $\tilde{a}^t$, where the q-th entry of $\tilde{a}^t$ is $\pi_q(\rho^t, t)$ if $q \in \mathcal{Y}$, and a^t otherwise. The safe policy $\tilde{\pi}$ for stage $t' \in \{t+1, \ldots, T-1\}$ can be computed as $\tilde{\pi} = \operatorname{argmin}_{\tilde{\pi}'} Q(s, \tilde{\pi}', T-t)$ for any joint state s.

Using the definitions of Q-function and safe policies, the probability of reaching the unsafe states in the counterfactual world $\mathcal{C}^t_{\mathcal{Y}}$ can be represented as

$$r(C^t_{\mathcal{Y}}; \{\pi_i\}_{i \in \mathcal{Y}}, \tilde{\pi}) = Q(\rho^t, \tilde{a}^t, T-t), \tag{8}$$

where $\tilde{a}^t = (\tilde{a}^t_1, \ldots, \tilde{a}^t_N)$ is the joint action taken by all agents in counterfactual world $\mathcal{C}^t_{\mathcal{Y}}$ when joint state is ρ^t at stage t.

V. Efficient Computation of DoR

As the number of agents in the MMDP increases, the computational complexity of determining DoRs grows exponentially, rendering it intractable for large-scale multi-agent CPS. In particular, we identify two key factors contributing to the high computational complexity of calculating DoR.

The first factor that increases the computational complexity of calculating DoR is that Eq. (4) requires evaluating all possible permutations of agents in $\mathcal{Y}$. Furthermore, for each permutation, the value of ϕ_i needs to be computed for each individual agent i. The second factor is that the joint state and action spaces of the MMDP grow exponentially with the number of agents. Consequently, there are a total of $T \prod_{i=1}^{N} |\mathcal{S}_i| \prod_{i=1}^{N} |\mathcal{A}_i|$ number of Q-functions that need to be calculated in Eq. (6)-(7) in order to apply Eq. (8).

In what follows, we first develop a heuristic algorithm to efficiently compute DoRs for arbitrary MMDPs. We then demonstrate that DoRs can be computed efficiently when the MMDP $\mathbb{M}$ possesses special structural properties.

We defer the proofs of our results to an extended version due to space constraints.

A. Heuristic Algorithm for Efficient Computation of DoR

In this subsection, we mitigate the computational complexity of our approach by developing an algorithm that identifies a subset of agents that are likely to be responsible for the observed path ρ reaching the unsafe states $\hat{\mathcal{S}}$. We denote the subset of responsible agents as $\mathcal{R} \subseteq \mathcal{N}$. By identifying the set $\mathcal{R}$, the value of ϕ_i can then be calculated as

$$\phi_i = \frac{1}{|\mathcal{N}|} \sum_{\mathcal{Y} \subseteq \mathcal{R} \setminus \{i\}} \binom{\mathcal{R}}{|\mathcal{R}| - |\mathcal{Y}| - 1}^{-1} (u(\mathcal{Y} \cup \{i\}) - u(\mathcal{Y})), \tag{9}$$

which significantly reduces the number of permutations that need to be evaluated, i.e., $2^{|\mathcal{R}|-1}$ coalitions in Eq. (9) compared with $2^{|\mathcal{N}|-1}$ coalitions in Eq. (4). In the worst-case where $\mathcal{R} = \mathcal{N}$, the computational complexity of evaluating ϕ_i using Eq. (9) is the same as using Eq. (4).

The key insight enabling identification of responsible agents is development of a concept named ϵ-marginally safe policy.

Definition 2 (ϵ-Marginally Safe Policy)**.** *Given a joint state s^t and joint action $a^t = (a_1^t, \ldots, a_N^t)$ at stage t, a policy π_i is said to be an ϵ-marginally safe policy for agent i if*

$$Q(s^t, a^t, T-t) - Q(s^t, \hat{a}^t, T-t) \geq \epsilon, \tag{10}$$

where $\hat{a}^t = (a_1^t, \ldots, a_{i-1}^t, \pi_i(s^t), a_{i+1}^t, \ldots, a_N^t)$.

We use Definition 2 and identify the set of responsible agents in Algorithm 1. The algorithm takes MMDP $\mathbb{M}$, observed path $\rho = s^0, a^0, s^1, \ldots, a^{T-2}, s^{T-1}$, and a threshold ϵ as inputs. It first initializes the set of responsible agents as $\mathcal{R} = \emptyset$, and then iterates over all agents $i \in \mathcal{N}$. At each iteration, the algorithm computes an ϵ-marginally safe policy for each agent i, assuming that other agents follow the same actions a_j^t at state ρ^t for all $j \in \mathcal{N} \setminus \{i\}$. Existence of an ϵ-marginally safe policy for an agent i at stage t can be verified by searching for an action $\hat{a}_i^t$ such that Eq. (10) holds. In the worst-case, computational complexity of searching for an ϵ-marginally safe policy of agent i is $\mathcal{O}(|\mathcal{A}_i|)$ at each stage t. Therefore, the computational complexity of Algorithm 1 is $\max_{i \in \mathcal{N}} TN\mathcal{O}(|\mathcal{A}_i|)$, which is linear in number of agents.

Algorithm 1 Identification of Responsible Agents $\mathcal{R}$

```
Input: MMDP 𝕄, observed path ρ, threshold ε
Output: The set of responsible agents ℛ ⊆ 𝒩
Initialize: ℛ ← ∅
for i ∈ 𝒩 do
    for t ∈ {0, ..., T − 1} do
        Let each agent j ∈ 𝒩 \ {i} take the action a_j^t
        Search for an action â_i^t ≠ a_i^t for agent i such that Eq. (10) holds
        if there exists such an action â_i^t then
            ℛ ← ℛ ∪ {i}
        end if
    end for
end for
Return ℛ
```

B. Structural MMDP Property for Efficient DoR Computation

We demonstrate that when the MMDP $\mathbb{M}$ satisfies certain structural properties, the DoRs can be calculated efficiently.

In the remainder of this section, we assume that interactions among agents are governed by an underlying undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes representing agents, and $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ is a set of edges. We denote the set of neighboring agents of an agent i as $\mathcal{B}_i$, i.e., for any agent $j \in \mathcal{B}_i$, there exists an edge $e_{ij} \in \mathcal{E}$ connecting agents i and j. The k-hop neighbors of an agent i is denoted as $\mathcal{B}_i^k$. We also define the states of the agents within and without the k-hop neighbors of agent i as $s_i^{\mathcal{B}^k}$ and $s_{-i}^{\mathcal{B}^k}$, respectively. Actions of the agents within and without the k-hop neighbors of agent i are denoted similarly as $a_i^{\mathcal{B}^k}$ and $a_{-i}^{\mathcal{B}^k}$, respectively. Therefore, for any joint state $s \in \mathcal{S}$ and joint action $a \in \mathcal{A}$, they can be rewritten as $s = (s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^k})$ and $a = (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k})$. For any joint state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, we assume that the transition probability Pr in Definition 1 can be represented as [23]

$$Pr(s'|s,a) = \prod_{i \in \mathcal{N}} Pr_i(s_i'|s_i^{\mathcal{B}}, a_i), \tag{11}$$

where $s_i^{\mathcal{B}}$ represents the joint states of agent i and its neighboring agents, and Pr_i is the transition function of agent i.

In what follows, we assume that the MMDP $\mathbb{M}$ satisfies an exponential decay property, which is defined below.

Definition 3 ((c,γ) Exponential Decay Property [23])**.** *The (c,γ) exponential decay property holds for some $c > 0$ and $\gamma \in (0,1)$ if the following inequality holds*

$$\Big| Q\left((s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^k}), (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k}), 0\right) - Q\left((s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^{k'}}), (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^{k'}}), 0\right) \Big| \leq c\gamma^{k+1} \tag{12}$$

for any agent $i \in \mathcal{N}$, joint states $(s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^k}), (s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^{k'}}) \in \mathcal{S}$, and joint actions $(a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k}), (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^{k'}}) \in \mathcal{A}$.

MMDPs that satisfy the (c,γ) exponential decay property are commonly encountered in practical scenarios. For example, when a set of vehicles collides with an obstacle, it is improbable that a vehicle located far away from the collision site will be responsible for the safety violation. This is because impact of maneuvers executed by the distant vehicle diminishes rapidly as it is spatially separated from the collision, consistent with the (c,γ) exponential decay property.

Given Definition 3, we can approximate the Q-function using a local Q-function, denoted as Q^L, as follows:

$$\begin{aligned} Q^L(s_i^{\mathcal{B}^k}, a_i^{\mathcal{B}^k}, 0) &= \sum_{s_{-i}^{\mathcal{B}^k} \in \mathcal{S}_{-i}^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k} \in \mathcal{A}_{-i}^{\mathcal{B}^k}} \omega_i\left((s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^k}), (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k})\right) \\ &\quad \cdot Q\left((s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^k}), (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k}), 0\right), \end{aligned} \tag{13}$$

where $\omega_i\left((s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^k}), (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k})\right)$ are non-negative weights such that

$$\sum_{s_{-i}^{\mathcal{B}^k} \in \mathcal{S}_{-i}^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k} \in \mathcal{A}_{-i}^{\mathcal{B}^k}} \omega_i\left((s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^k}), (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k})\right) = 1, \tag{14}$$

$\mathcal{S}_{-i}^{\mathcal{B}^k}$ is the set of states of all agents outside the k-hop neighborhood of agent i, and $\mathcal{A}_{-i}^{\mathcal{B}^k}$ is the set of actions of all agents outside the k-hop neighborhood of agent i. We have the following optimality guarantee when approximating the Q-function using the local Q-function in Eq. (13).

Lemma 1. *Let $s = (s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^{k'}}) \in \mathcal{S}$ and $a = (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^{k'}}) \in \mathcal{A}$ be the joint state and action of the MMDP $\mathbb{M}$. If the MMDP $\mathbb{M}$ satisfies the (c,γ) exponential decay property in Definition 3, then the following relationship holds*

$$\sup_{s_{-i}^{\mathcal{B}^{k'}} \in \mathcal{S}_{-i}^{\mathcal{B}^{k'}}, a_{-i}^{\mathcal{B}^{k'}} \in \mathcal{A}_{-i}^{\mathcal{B}^k}} |Q^L(s_i^{\mathcal{B}^k}, a_i^{\mathcal{B}^k}, 0) - Q(s,a,0)| \leq c\gamma^{k+1}.$$

Lemma 1 shows that we can approximate the 0-stage-to-go Q-function using only the local Q-function, which requires tracking joint states and actions of a subset of agents, i.e., agents within the k-hop neighborhood of agent i. By Lemma 1, we can approximate the Q-function using local Q-functions for arbitrary $t \in \{1, \ldots, T-1\}$, and thus approximate safe policies for each agent in a distributed manner as follows:

$$Q^L(s_i^{\mathcal{B}^k}, a_i^{\mathcal{B}^k}, t) = \sum_{s_i^{\mathcal{B}^{k'}} \in \mathcal{S}_i^{\mathcal{B}^k}} Pr^L(s_i^{\mathcal{B}^{k'}} | s_i^{\mathcal{B}^k}, a_i^{\mathcal{B}^k}) \cdot \min_{a_i^{\mathcal{B}^{k'}}} Q^L(s_i^{\mathcal{B}^k}, a_i^{\mathcal{B}^{k'}}, t-1), \ \forall t \in \{1, \ldots, T-1\}, \quad (15)$$

where $Pr^L(s_i^{\mathcal{B}^{k'}} | s_i^{\mathcal{B}^k}, a_i^{\mathcal{B}^k})$ is the marginalized transition probability given as

$$Pr^L\left(s_i^{\mathcal{B}^{k'}} | s_i^{\mathcal{B}^k}, a_i^{\mathcal{B}^k}\right) = \sum_{\substack{s_{-i}^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^{k'}} \in \mathcal{S}_{-i}^{\mathcal{B}^k} \\ a_{-i}^{\mathcal{B}^k} \in \mathcal{A}_{-i}^{\mathcal{B}^k}}} Pr\left((s_i^{\mathcal{B}^{k'}}, s_{-i}^{\mathcal{B}^{k'}}) | (s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^k}), (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^k})\right). \quad (16)$$

The optimality guarantee of the local Q-function obtained in Eq. (15) is given in the following result.

Proposition 2. *Let $s = (s_i^{\mathcal{B}^k}, s_{-i}^{\mathcal{B}^{k'}}) \in \mathcal{S}$ and $a = (a_i^{\mathcal{B}^k}, a_{-i}^{\mathcal{B}^{k'}}) \in \mathcal{A}$ be the joint state and action of the MMDP $\mathbb{M}$. If the MMDP $\mathbb{M}$ satisfies the (c, γ) exponential decay property in Definition 3, then for any stage $t = 0, \ldots, T-1$, the following relationship holds*

$$\sup_{s_{-i}^{\mathcal{B}^{k'}} \in \mathcal{S}_{-i}^{\mathcal{B}^{k'}}, a_{-i}^{\mathcal{B}^{k'}} \in \mathcal{A}_{-i}^{\mathcal{B}^k}} |Q^L(s_i^{\mathcal{B}^k}, a_i^{\mathcal{B}^k}, t) - Q(s, a, t)| \leq c\gamma^{k+1}.$$

We note that we only need to evaluate the local Q-function on a smaller set of joint states and actions. This allows us to utilize much less memory to store the local Q-function and less computational power to calculate the approximately safe policies, making our solution technique more tractable for large-scale multi-agent CPS. We remark that the optimality bound in Lemma 1 decreases exponentially as the state and action of more agents are revealed. We finally present the optimality bound when approximating ϕ_i in Eq. (4) using the local Q-function as follows.

Theorem 1. *Assume that the MMDP $\mathbb{M}$ satisfies the (c, γ) exponential decay property in Definition 3. For an agent i, let ϕ_i be defined in Eq.* (4) *and evaluated using Eq.* (8)*. Let the probability of reaching the unsafe states in each counterfactual world $\mathcal{C}_{\mathcal{Y}}^t$ be approximated as*

$$r(C_{\mathcal{Y}}^t; \{\pi_i\}_{i \in \mathcal{Y}}, \tilde{\pi}) \approx Q^L(s_i^{\mathcal{B}^k}, a_i^{\mathcal{B}^k}, T-t),$$

and ϕ_i^L be the associated utility distributed to agent i within the coalition $\mathcal{Y}$. We have $|\phi_i - \phi_i^L| \leq c^L \gamma^{k+1}$ for all agent $i \in \mathcal{N}$, where

$$c^L = 2\frac{1}{|\mathcal{N}|} \sum_{\mathcal{Y} \subseteq \mathcal{N} \setminus \{i\}} \binom{\mathcal{N}}{|\mathcal{N}| - |\mathcal{Y}| - 1}^{-1} c.$$

Theorem 1 shows that if the MMDP $\mathbb{M}$ satisfies the (c, γ) exponential decay property, then we can use the local Q-function to approximate the safe policies and calculate the DoR for each agent. The approximation incurs a bounded error as shown in Theorem 1. Furthermore, the error decreases exponentially when the local Q-function utilizes the states from more agents to approximate the Q-function.

VI. Experiments

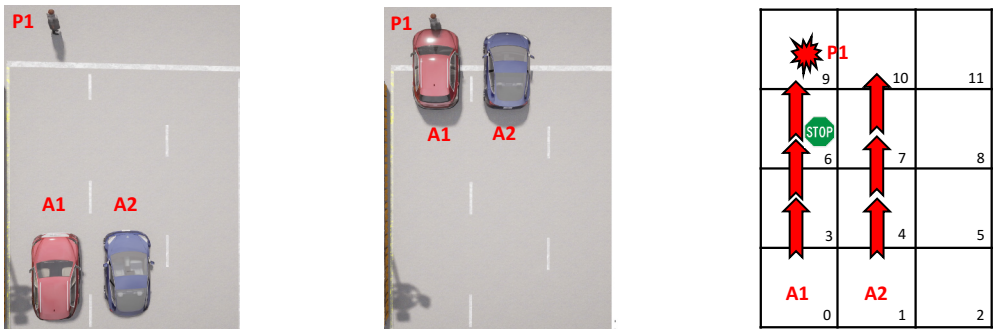
In this section, we consider three instances of safety violations (i.e., an accident or crash) available from the National Highway Traffic Safety Administration (NHTSA) [24]. Each instance includes information about vehicle movements and dynamics, as well as critical events immediately preceding a crash. We render these scenarios using CARLA, an open-source urban driving simulator [41].

We conducted our experiments using a workstation installed with the Linux 5.19.0-43-generic operating system. The workstation is equipped with an Intel® Xeon® W-2145 CPU @ 3.70 GHz processor, two 16 GB NVIDIA GeForce RTX 2080 Ti GPUs, and 128 GB of memory. We used cvxpy version 0.9.14 within a Python 3.7 environment to construct three road safety violation scenarios. We used the Python MDP Toolbox [42] to determine the safe policies for vehicle agents after modeling each scenario as an MMDP in Definition 1.

We first provide details about representation of the three scenarios using CARLA and describe how to construct the MMDP using data collected from CARLA for each scenarios. Next, we present our experimental results and discussions related to these simulated scenarios. Finally, we discuss some scenarios that are not addressed by our present DoR formulation, but nevertheless form a basis for future research.

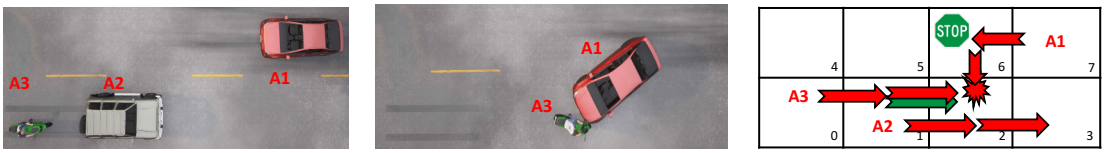
A. CARLA Simulated Scenarios

We demonstrate our approach using three road safety violation scenarios derived from [24], and simulated in CARLA. **Scenario 1:** Our first scenario consists of two vehicles and one non-moving pedestrian in a three-lane road segment. Vehicle Agents 1 and 2 drive parallel to each other and Agent 1 collides with the pedestrian. Snapshots of the initial locations and collision are illustrated in Fig. 2-(a) and-(b), respectively. **Scenario 2:** Our second scenario consists of three vehicles in a two-lane road segment. Vehicle Agents 2 and 3 operate in the same lane, and move in the same direction. Agent 1, which is operating in the opposite lane to Agents 2 and 3, takes a U-turn. When Agent 1 is completing the U-turn, it collides with Agent 3. Snapshots of the initial locations of all vehicles and the collision are presented in Fig. 3-(a) and-(b), respectively. **Scenario 3:** Our third scenario has two vehicles. Vehicle Agent 2 stays within its current lane on the highway. Vehicle Agent



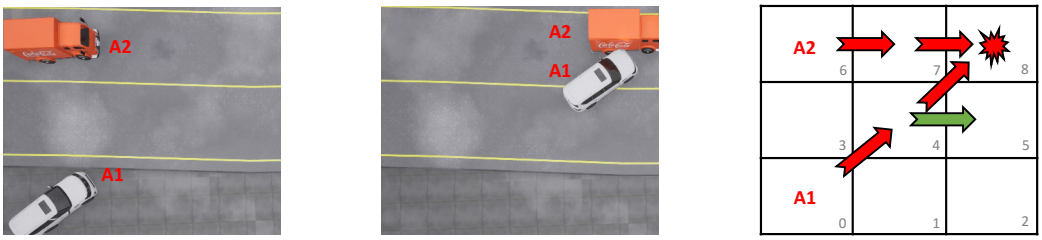
(a) Initial locations of vehicles (b) Illustration of the collision (c) Illustration of the discretized locations

Fig. 2: **Scenario 1**: This figure presents the initial locations of vehicles, the collision, and the discretized locations for Scenario 1. The red car is Agent 1 and the blue car is Agent 2. The initial locations of Agents 1 and 2 (labeled as A1 and A2) are presented in Fig. 2-(a). Fig. 2-(b) illustrates a collision, where Agent 1 collides with a pedestrian (P1). The road segment is discretized into 12 locations $\mathcal{L} = \{0, \ldots, 11\}$ as shown in Fig. 2-(c). Red arrows represent paths observed in path ρ, and the green STOP sign represents the safe policy of Agent 1. Here the STOP sign indicates that the agent should take the stop action to avoid safety violation.



(a) Initial locations of vehicles (b) Illustration of the collision (c) Illustration of the discretized locations

Fig. 3: **Scenario 2**: This figure presents the initial locations of vehicles, the collision, and the discretized locations for Scenario 2. The red car is Agent 1, the white SUV is Agent 2, and the green motorcycle is Agent 3. The initial locations of Agents 1, 2, and 3 (labeled as A1, A2, and A3) are presented in Fig. 3-(a). Fig. 3-(b) illustrates the collision, where Agent 1 collides with Agent 3. The road segment is discretized into 8 locations $\mathcal{L} = \{0, \ldots, 7\}$, as shown in Fig. 3-(c). Red arrows represent the paths observed in path ρ, and the green STOP sign represents the safe policy of Agents 1 and 3. Here the STOP sign indicates that the agent should take the stop action to avoid safety violation.



(a) Initial locations of vehicles (b) Illustration of the collision (c) Illustration of the discretized locations

Fig. 4: **Scenario 3**: This figure presents the initial locations of vehicles, the collision, and the discretized locations for Scenario 3. The white SUV is Agent 1, and the orange truck is Agent 2. The initial locations of Agents 1 and 2 (labeled as A1 and A2) are presented in Fig. 4-(a). Fig. 4-(b) illustrates the collision, where Agent 1 collides with Agent 2. The road segment is discretized into 9 locations $\mathcal{L} = \{0, \ldots, 8\}$ as shown in Fig. 4-(c). Red arrows represent the paths observed in path ρ, and the green arrow represents the safe policy of Agent 1.

1 departs from the gas station and aims to merge into the lane in which Agent 2 is present. Agents 1 and 2 collide with each other when Agent 1 tries to merge. Snapshots of initial locations and the collision are illustrated in Fig. 4-(a) and-(b).

MMDP Construction: We discuss how to construct an MMDP for each scenario described above. We first collect a video stream of length 10 seconds prior to the traffic accident from the log file available from the CARLA simulator. This video can then be used to obtain time-stamped trajectories of vehicles and status of all traffic lights. In real-world applications, video stream captured by RGB-D cameras can be used to estimate vehicle trajectories and signal status [43].

We then discretize the environment containing all trajectories into a set of discrete locations $\mathcal{L}$, where each location is represented by a cell. The length and width of each cell are chosen as the length of the vehicles and width of the lanes, respectively. Discretized locations for Scenarios 1, 2, and 3 are illustrated in Fig. 2-(c), Fig. 3-(c), and Fig. 4-(c). Then, the joint state space of the MMDP is constructed as $\mathcal{S} = \mathcal{L}^{|\mathcal{N}|}$. Each agent has at most four actions- moving forward straight ahead (move forward), forward left, forward right, and stop. Some actions will be forbidden at certain states, e.g., moving forward left or right is not allowed if there is only one lane. We focus on the best-case where agents can always successfully transition among the locations, i.e., transition probability to desired location is one when the vehicle takes desired action.

Scenario ID	Agent ID	DoR	Memory	Runtime
1	1	1	6.69 Mb	3.00 sec
	2	0		
2	1	0.5	115.29 Mb	88.49 sec
	2	0		
	3	0.5		
3	1	1	5.43 Mb	1.55 sec
	2	0		

TABLE I: Values of Degree of Responsibility (DoR) assigned to vehicle agents following our proposed approach for experiments in CARLA for scenarios in Fig. 2, Fig. 3, and Fig. 4. The table additionally shows memory usage and runtime to compute the DoRs in each case.

B. Experiment Results

Table I presents DoR values for each vehicle agent in the three road safety violation scenarios shown in Fig. 2 to Fig. 4 along with memory usage and computational runtime.

Our method assigns DoR of 1 and 0 to Agents 1 and 2 in Scenario 1. This is because Agent 1 could have acted to stop the vehicle and avoid colliding with the pedestrian, regardless of Agent 2's action. For e.g., Agent 1 could have followed a safe policy, yielding path $(0, \text{move forward}, 3, \text{move forward}, 6, \text{stop}, 6)$ to avoid the pedestrian in location 9, even when Agent 2 follows the same path (1, move forward, 4, move forward, 7, move forward, 10) as observed in ρ. Therefore, Agent 1 is fully responsible (i.e., DoR $= 1$), whereas Agent 2 bears no responsibility.

For Scenario 2, Agents 1 and 3 share responsibility while Agent 2 bears no responsibility for the collision. An explanation for these DoR values is that Agent 2 can safely cruise through the road segment by following same actions observed in path ρ regardless of actions of Agents 1 and 3. However, actions of Agents 1 and 3 cannot guarantee that the safety constraint will be satisfied. Instead, Agents 1 and 3 will need to cooperate to avoid collision. For e.g., when Agent 3 follows actions observed in ρ, safety can be guaranteed when Agent 1 follows the safe policy (reach location 6 and then stop until Agent 3 reaches location 3). Similarly, Agent 1 can follow the safe policy even when Agent 3 takes the action observed in ρ_3. Thus, Agents 1 and 3 each have DoR $= 0.5$.

Our approach asserts that Agent 1 is fully responsible in Scenario 3. Here, Agent 2 does not have alternate actions available in any counterfactual world except observed actions, i.e., cruising on the highway in the current lane. The insight underpinning our DoR calculation is that Agent 1 (white SUV) could have taken some alternative actions (e.g., merge into rightmost lane, or stop before merging into highway) to avoid colliding with Agent 2 (orange truck). As a result, Agent 1 is assigned $DoR = 1$ while Agent 2 is assigned $DoR = 0$.

C. Discussion

The notion of DoR developed in this paper and the three scenarios above describe the impact and analysis following a single crash. In order to investigate a situation such as a cascade collision of vehicles resulting from the abrupt braking of a leading vehicle will necessitate reasoning about and formulating compositional properties of the DoR. Our DoR also does not explicitly account for the type of vehicle (e.g., car, truck) that is involved in an accident. Such compositional reasoning may also inform assigning levels of responsibility when multiple types of vehicles are involved in an accident.

Our formulation of DoR solely focuses on quantifying responsibility arising from violations of safety due to erroneous decision-making in multi-agent CPS. In practice, there are several other factors that can contribute to ascribing degrees of responsibility to agents involved in an accident. Examples of these factors include considerations of conventional norms and the letter of the law [44]. Examining the interplay among safety, legal, and conventional norms and integrating these into a unified quantitative score to assign responsibility following a safety violation will lead to a richer and more interpretable variant of DoR.

VII. Conclusion

This paper developed a methodology to identify and quantify responsibility to agents in a multi-agent cyber-physical system when errors in agent decision making led to safety violations. The design of a principled automated procedure to assign responsibility was informed by a need to reduce reliance on human expertise and resulting cognitive burden associated with current methodologies. The responsibility of an individual agent relative to other agents was quantified using a metric called the degree of responsibility (DoR) to provide an auditable interpretation of agent responsibility using counterfactual reasoning. We considered three instances

of safety violations from the National Highway Traffic Safety Administration (NHTSA) and performed extensive experiments using representations of the three scenarios in the CARLA urban driving simulator.

We hypothesize that our DoR metric can also complement explanation-based techniques focused on reasoning about machine learning models [38]. Future work will provide complete proofs of our theoretical results, and expand empirical validation through additional case studies.

REFERENCES

[1] J. Van den Berg, M. Lin, and D. Manocha, "Reciprocal velocity obstacles for real-time multi-agent navigation," in *IEEE International Conference on Robotics and Automation*. IEEE, 2008, pp. 1928–1935.

[2] J. Alonso-Mora, A. Breitenmoser, M. Rufli, P. Beardsley, and R. Siegwart, "Optimal reciprocal collision avoidance for multiple non-holonomic robots," in *Distributed Autonomous Robotic Systems: The 10th International Symposium*. Springer, 2013, pp. 203–216.

[3] U. Borrmann, L. Wang, A. D. Ames, and M. Egerstedt, "Control barrier certificates for safe swarm behavior," *IFAC-PapersOnLine*, vol. 48, no. 27, pp. 68–73, 2015.

[4] J. Chen, J. Li, C. Fan, and B. C. Williams, "Scalable and safe multi-agent motion planning with nonlinear dynamics and bounded disturbances," in *AAAI Conference on Artificial Intelligence*, 2021.

[5] L. Wang, A. D. Ames, and M. Egerstedt, "Safety barrier certificates for collisions-free multirobot systems," *IEEE Transactions on Robotics*, vol. 33, no. 3, pp. 661–674, 2017.

[6] Y. Chen, A. Singletary, and A. D. Ames, "Guaranteed obstacle avoidance for multi-robot operations with limited actuation: A control barrier function approach," *IEEE Control Systems Letters*, vol. 5, no. 1, pp. 127–132, 2020.

[7] E. Yel, T. J. Carpenter, C. Di Franco, R. Ivanov, Y. Kantaros, I. Lee, J. Weimer, and N. Bezzo, "Assured runtime monitoring and planning: Toward verification of neural networks for safe autonomous operations," *IEEE Robotics & Automation Magazine*, vol. 27, no. 2, 2020.

[8] A. Lavaei, S. Soudjani, and E. Frazzoli, "A compositional dissipativity approach for data-driven safety verification of large-scale dynamical systems," *IEEE Transactions on Automatic Control*, 2023.

[9] T. Chu, S. Chinchali, and S. Katti, "Multi-agent reinforcement learning for networked system control," *International Conference on Learning Representations*, 2020.

[10] S. Narasimhan, S. Bhat, and S. Chinchali, "Safe networked robotics with probabilistic verification," *IEEE Robotics & Automation Letters*, 2023.

[11] F. Cai and X. Koutsoukos, "Real-time out-of-distribution detection in learning-enabled cyber-physical systems," in *ACM/IEEE 11th International Conference on Cyber-Physical Systems*, 2020, pp. 174–183.

[12] Y. Yang, R. Kaur, S. Dutta, and I. Lee, "Interpretable detection of distribution shifts in learning enabled cyber-physical systems," in *ACM/IEEE International Conf. on Cyber-Physical Systems*, 2022, pp. 225–235.

[13] W. H. Organization, "Road traffic injuries," https://www.who.int/en/news-room/fact-sheets/detail/road-traffic-injuries.

[14] S. of California Department of Motor Vehicles, "Autonomous vehicle collision reports," https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/.

[15] H. A. Yawovi, M. Kikuchi, and T. Ozono, "Who was wrong? An object detection based responsibility assessment system for crossroad vehicle collisions," *AI*, vol. 3, no. 4, pp. 844–862, 2022.

[16] A. F. Winfield, K. Winkle, H. Webb, U. Lyngs, M. Jirotka, and C. Macrae, "Robot accident investigation: A case study in responsible robotics," *Software Engineering for Robotics*, pp. 165–187, 2021.

[17] D. Omeiza, H. Web, M. Jirotka, and L. Kunze, "Towards accountability: Providing intelligible explanations in autonomous driving," in *IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 231–237.

[18] Đ. Petrović, R. Mijailović, and D. Pešić, "Automated vehicle data pipeline for accident reconstruction: New insights from LiDAR, camera, and radar data," *Accident Analysis & Prevention*, vol. 180, 2023.

[19] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson, "Counterfactual reasoning and learning systems: The example of computational advertising." *Journal of Machine Learning Research*, vol. 14, 2013.

[20] K. Epstude and N. J. Roese, "The functional theory of counterfactual thinking," *Personality and Social Psychology Review*, vol. 12, no. 2, pp. 168–192, 2008.

[21] A. E. Roth, *The Shapley value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[22] J. Y. Halpern and J. Pearl, "Causes and explanations: A structural-model approach. part i: Causes," *The British journal for the philosophy of science*, 2005.

[23] G. Qu, Y. Lin, A. Wierman, and N. Li, "Scalable multi-agent reinforcement learning for networked systems with average reward," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2074–2086, 2020.

[24] N. H. T. S. Administration, "Pre-crash scenario typology for crash avoidance research," https://www.nhtsa.gov/sites/nhtsa.gov/files/pre-crash_scenario_typology-final_pdf_version_5-2-07.pdf.

[25] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.

[26] H. Chockler and J. Y. Halpern, "Responsibility and blame: A structural-model approach," *Journal of Artificial Intelligence Research*, vol. 22, pp. 93–115, 2004.

[27] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," *arXiv preprint arXiv:1610.03295*, 2016.

[28] R. Cheng, M. J. Khojasteh, A. D. Ames, and J. W. Burdick, "Safe multi-agent interaction through robust control barrier functions with learned uncertainties," in *IEEE Conf. Decision and Control*, 2020, pp. 777–783.

[29] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 1343–1350.

[30] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in Neural Information Processing Systems*, 2017.

[31] M. Everett, Y. F. Chen, and J. P. How, "Motion planning among dynamic, decision-making agents with deep reinforcement learning," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2018, pp. 3052–3059.

[32] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5872–5881.

[33] Z. Qin, K. Zhang, Y. Chen, J. Chen, and C. Fan, "Learning safe multi-agent control with decentralized neural barrier certificates," in *International Conference on Learning Representations*, 2021.

[34] I. ElSayed-Aly, S. Bharadwaj, C. Amato, R. Ehlers, U. Topcu, and L. Feng, "Safe multi-agent reinforcement learning via shielding," in *Int. Conf. on Autonomous Agents and MultiAgent Systems*, 2021, p. 483–491.

[35] J. Torres, "Investigating traffic crashes involving autonomous vehicles," in *IISE Annual Conference Proceedings*, 2021.

[36] Đorđe Petrović, R. Mijailović, and D. Pešić, "Traffic accidents with autonomous vehicles: Type of collisions, manoeuvres and errors of conventional vehicles' drivers," *Transportation Research Procedia*, vol. 45, pp. 161–168, 2020.

[37] S. Lee, R. Arvin, and A. J. Khattak, "Advancing investigation of automated vehicle crashes using text analytics of crash narratives and Bayesian analysis," *Accident Analysis & Prevention*, vol. 181, 2023.

[38] B. Rozemberczki, L. Watson, P. Bayer, H.-T. Yang, O. Kiss, S. Nilsson, and R. Sarkar, "The shapley value in machine learning," in *International Joint Conference on Artificial Intelligence*, 2022, pp. 5572–5579.

[39] D. Fudenberg and J. Tirole, *Game Theory*. MIT press, 1991.

[40] S. Han, H. Wang, S. Su, Y. Shi, and F. Miao, "Stable and efficient Shapley value-based reward reallocation for multi-agent reinforcement learning of autonomous vehicles," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8765–8771.

[41] 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.

[42] C. Heins, B. Millidge, D. Demekas, B. Klein, K. Friston, I. D. Couzin, and A. Tschantz, "pymdp: A Python library for active inference in discrete state spaces," *Journal of Open Source Software*, vol. 7, no. 73, p. 4098, 2022.

[43] Z. Shan, Q. Zhu, and D. Zhao, "Vehicle collision risk estimation based on RGB-D camera for urban road," *Multimedia systems*, vol. 23, pp. 119–127, 2017.

[44] T. M. Rudisill and M. Zhu, "Hand-held cell phone use while driving legislation and observed driver behavior among population sub-groups in the united states," *BMC public health*, vol. 17, pp. 1–10, 2017.