Towards the Unification and Data-Driven Synthesis of Autonomous Vehicle Safety Concepts

Anonymous Author(s)

Affiliation
Address
e-mail

Abstract: As safety-critical autonomous vehicles (AVs) will soon become pervasive in our society, a number of safety concepts for trusted AV deployment have been recently proposed throughout industry and academia. Yet, agreeing upon an “appropriate” safety concept is still an elusive task. In this paper, we advocate for the use of Hamilton Jacobi (HJ) reachability as a unifying mathematical framework for comparing existing safety concepts, and propose ways to expand its modeling premises in a data-driven fashion. Specifically, we show that (i) existing predominant safety concepts can be embedded in the HJ reachability framework, thereby enabling a common language for comparing and contrasting modeling assumptions, and (ii) HJ reachability can serve as an inductive bias to effectively reason, in a data-driven context, about two critical, yet often overlooked aspects of safety: responsibility and context-dependency.

Keywords: safety, autonomous vehicles, HJ reachability

1 Introduction

A rigorous and actionable understanding of safety is an essential prerequisite for autonomous vehicle (AV) deployment and widespread adoption. To this end, stakeholders including AV industry players, academics, and policymakers have advanced various safety concepts for evaluating driving scenarios and deciding what constitutes safe behavior. Underpinning these notions of safety are modeling assumptions reasoned from first principles or informed by data. However, to date there is no universal agreement on how to select these assumptions and therefore how safety should be assessed and embedded in modern AVs. In fact, investigating such assumptions is an active topic in the standards community [1]. In this paper, we do not purport to resolve which proposed safety concept, if any, is best. Instead, we aim to accelerate convergence to a shared consensus by casting these concepts within a common framework to promote comparison and novel synthesis.

In this work, we define safety concepts as the combination of two functions mapping world state (e.g., joint state of all agents and environmental context like road geometry) to (i) a scalar measure of safety, and (ii) a set of allowable actions for each agent. We note that this definition abstracts away the computation by which modeling assumptions are transformed into such functions; indeed, factors such as measurement uncertainty, delayed reaction times, and possibly stochastic models of behavior are considered implicitly. Aside from clear-cut cases where, for example, inevitable collision may be determined purely from dynamics considerations, deciding on a safety concept is challenging because of the necessary dependence on responsibility (degree of each agent’s ownership over collision-avoidance) as well as context. In particular, safety and responsibility are inextricably entwined, whereby the safety of all but the most conservative maneuvers relies on the assumption that surrounding agents will act responsibly, while even defining responsibility requires a means to quantify how much agents’ actions contribute to safety. The nuance of how safety and responsibility are
influenced by context (e.g., a road blockage may allow an agent to briefly cross into an oncoming traffic lane which would otherwise be wholly irresponsible) motivates the usage of driving data in learning safety concepts. How to do this in a principled fashion, however, remains an open research question.

We propose that Hamilton-Jacobi (HJ) reachability is a promising unifying mathematical framework for describing safety concepts and provides, by design, the inductive bias that the scalar safety measure and allowable action sets should be consistent. In this work, we take the first steps to demonstrate that many seemingly disparate existing safety concepts can be unified via HJ reachability. We note that HJ reachability on its own does not answer the important question of which safety concept is most appropriate for AVs. Instead, it enables us to characterize a family of safety concepts for which we may hope to design data-driven synthesis techniques for novel safety concepts. We further elaborate research directions on safety-centric dataset construction, interpretable representations of context, and learned notions of responsibility, highlighting the role of HJ reachability as inductive bias throughout. Our goal is to help stakeholders converge on a safety concept grounded in data, reflecting both realistic AV interactions while remaining interpretable from a policy perspective.

2 A unifying framework for safe autonomous vehicle interactions

In this section, we provide a mathematical introduction to HJ reachability while highlighting key elements that make it possible to encompass a family of safety concepts, and provide valuable inductive bias during data-driven synthesis and evaluation of safety concepts. By demonstrating the modeling flexibility of HJ reachability, we can more confidently take the first steps towards novel safety concepts which harness data to reason about responsibility and context-dependency.

2.1 Background: Hamilton-Jacobi Reachability

HJ reachability is a mathematical formalism for characterizing the performance and safety properties of (multi-agent) dynamical systems [2, 3, 4]. Core to HJ reachability is the set of states \( \mathcal{L} \) which agents reason about either seeking or avoiding within a time horizon \( T \). In the context of collision avoidance between agents \( A \) and \( B \), \( \mathcal{L} \) corresponds to the set of collision states. To capture the multi-agent, continuous-time, and safety-critical nature of our setting, HJ reachability describes a two-player\(^1\) differential game. This formulation enables us to mathematically characterize whether it is possible for the ego agent to prevent an undesirable outcome under any family of closed-loop policies of other agents, as well as the ego agent’s appropriate control policy for ensuring safety.

Using the principle of dynamic programming, the collision avoidance problem between agents \( A \) and \( B \) reduces to solving the Hamilton-Jacobi-Isaacs (HJI) partial differential equation (PDE),

\[
\frac{\partial V(z, t)}{\partial t} + \min \left\{ 0, \max_{u_A \in \mathcal{U}_A} \min_{u_B \in \mathcal{U}_B} \nabla_z V(z, t)^\top f(z, u_A, u_B) \right\} = 0
\]

(1)

where \( z \in \mathcal{Z} \) denotes the joint state of agents \( A \) and \( B \), \( u_A \) and \( u_B \) are the controls of agents \( A \) and \( B \), respectively, and \( f(\cdot, \cdot, \cdot) \) is the joint dynamics. The boundary condition for this PDE is defined by the function \( \ell(z) \) whose zero sub-level set encodes the undesirable states \( \mathcal{L} \), i.e., \( \mathcal{L} = \{ z \mid \ell(z) < 0 \} \). Lastly, a more general reach-avoid formulation also exists, whereby the agents optimize with respect to both a goal set and avoid set. We refer the reader to [3, 5] for more details.

By solving Equation (1) backwards in time over a time horizon of \( T \), we obtain the HJ value function \( V(z, t) \) for \( t \in [-T, 0] \). For any starting state \( z \in \mathcal{Z} \), this function captures the closest the overall system can get to the set of undesirable states (i.e., the lowest value of \( \ell(\cdot) \) along a system trajectory) within \( |t| \) seconds if both agents \( A \) and \( B \) act optimally, \( u_A^*(z), u_B^*(z) = \max_{u_A \in \mathcal{U}_A} \min_{u_B \in \mathcal{U}_B} \nabla_z V(z, t)^\top f(z, u_A, u_B) \). Since (1) encodes a collision-avoidance problem, we can obtain the unsafe set \( \mathcal{A}(t) = \{ z \mid V(z, t) < 0 \} \). In short, the HJ value function encodes both the unsafe set and a scalar measure capturing the degree of safety-violation the system can encounter in the future.

\(^1\)In general, \( n \)-player. In this work, we focus on the two-player setting for brevity and tractability.
HJ reachability thus synthesizes a safety measure through the value function $V$, which fulfills the first aspect of a safety concept. Importantly, this framework also facilitates learning the second aspect—the allowable control sets—by ensuring consistency (in terms of dynamics and agent intent) between the control sets and their effect on the safety measure. The family of safety concepts that can be encoded and learned in Equation (1) ultimately depend on three key aspects of the HJ reachability formulation: (i) the type of agent reactions, (ii) allowable control sets, and (iii) the safety criterion.

**Type of agents’ reactions** (max or min): As currently stated in (1), agent $B$ has informational advantage (by “playing second” and responding to $A$’s controls) and agent $B$ is modelled as acting **adversarially** by minimizing the instantaneous rate of change of the value function subject to the joint dynamics. By modifying the information pattern and/or if agents are minimizers or maximizers, we can modify the strategic optimism when computing the unsafe set (Figure 2, left).

**Allowable control sets ($\mathcal{U}^A$ and $\mathcal{U}^B$):** Traditionally, the control set represents dynamically feasible controls of each agent (e.g., respecting actuation limits of the system). However, we can restrict the control set (e.g., informed by data) to reflect assumptions about how other agents behave in safety-critical scenarios (Figure 2, center).

**Safety criterion $\mathcal{L} = \{z \mid \ell(z) < 0\}$:** We are free to define $\ell(\cdot)$ as long as its zero sub-level set equals $\mathcal{L}$. While purely geometric functions like penetration/separation distance are a common choice, we can design or learn alternative functions that capture more nuanced notions of safety. For example, by shaping $\ell$ to penalize more dangerous orientations (e.g., T-bone or head-on collision) we can encode collision severity or collision responsibility (Figure 2, right).

![Figure 2: Elements of HJ reachability allowing a spectrum of safety concepts in the same framework.](image)

### 2.2 Embedding existing safety concepts in HJ reachability

Here, we briefly describe existing predominant safety concepts and demonstrate how they can be expressed via HJ reachability. We note that the majority of these concepts were developed without HJ reachability in mind. Throughout, we assume $\ell(\cdot)$ encodes penetration/separation distance.

**Worst-case dynamic game** [2, 6]: An **unsafe** state is one where a collision is inevitable despite agent $A$’s best effort to avoid collision and agent $B$’s policy to cause collision. **HJ formulation**: This worst-case **max min** formulation is stated in (1), allowing for all dynamically-feasible controls.

**Safety Force Field (SFF)** [7]: Assume each agent has a set of safety procedures, i.e., a set of control signals that brings an agent to a stop in finite time. An **unsafe** state is one where there exists a pair of safety procedures which, if executed, lead to agents A and B colliding. **HJ formulation**: This is a **min min** game (since we want to detect a pair of safety procedures that lead to a collision) where both agents optimize over a restricted control set containing only the safety procedures.

**Responsibility-Sensitive Safety (RSS)** [8]: An **unsafe** state is one where a collision occurs if both agents apply a fixed deceleration in both the longitudinal and lateral dimension. **HJ formulation**: Since longitudinal and lateral motions are decoupled, the corresponding unsafe set is the intersection of the longitudinal and lateral unsafe sets. To compute the longitudinal (lateral) unsafe set, we consider the longitudinal (lateral) dynamics and restrict the control set to only deceleration, i.e., $|\mathcal{U}^A| = 1$, $|\mathcal{U}^B| = 1$. With singleton control sets, the **max min** operations can be removed.

**Contingency safety** [9, 10, 11]: An **unsafe** state is one where an agent is not able to come to a collision-free stop assuming other agents maintain their current assumed motion. The notion

\(^2\)For brevity, we ignore reaction time, but can account for this with a time-varying control set.
of “contingency safety” describes a class of safety concepts, depending on how other agents are
assumed to behave. **HJ formulation:** We may set up a reach-avoid game where agent $A$ strives to
avoid agent $B$ while $A$ tries to reach a zero-velocity state. The type of game and agent $B$’s control
set depends on the assumptions on its behavior (e.g., maintain speed in lane vs. hard brake).

**Constant motion** [12, 13]: An unsafe state is one where if both agents maintain their current ve-
locity and steering, they will collide at some future time. **HJ formulation:** Augment the state-space
with the control inputs and set control derivatives to zero. As such, this state augmentation removes
the game-theoretic aspect of the HJ formulation.

### 3 Grounding responsibility and context-aware safety concepts in data

The above safety concepts largely lack nuanced treatments of responsibility (instead rigidly assum-
ing, for example, that agents follow prescribed policies in safety critical situations), and lack expres-
siveness in incorporating local context. In this section we argue for the necessity of such considerations
in deriving an ideal safety concept, and propose research directions according to the paradigm that
context-dependency and the coupling between responsibility and safety should be grounded in data.

**Data.** What data is necessary to learn safety concepts? This question requires embarking on two re-
search directions. First, [D1] developing new—or augmenting existing—datasets by actively query-
ing humans for safety-relevant labels. We contend that logged trajectory data alone may not capture
counterfactual-dependent questions such as how safe is a scenario?, did the agent(s) perform accept-
able maneuvers?, and, if a maneuver is labeled unsafe, what alternative(s) are preferable? Second,
[D2] datasets should ensure sufficient coverage of safety-critical scenarios while respecting a lim-
ited labeling budget. Here, policymakers can inform dataset coverage by enumerating a list of key
AV scenario classes, while engineering stakeholders can reduce human effort through automated
and targeted dataset augmentation; for example, by optimizing out-of-distribution metrics to capture
epistemic uncertainty or taking inspiration from the Quality-Diversity optimization community [14].

**Context.** Contextual reasoning has proven instrumental in advancing understanding in AV-relevant
fields, e.g., as enabled in computer vision by the Common Objects in Context (COCO) dataset
[15]. Towards capturing richer notions of AV safety without losing first-principles interpretability,
we advocate for [D3] establishing representations of context that are tangible for stakeholders and
operationalizable within safety computation. Incorporating context will be impactful in cases where
even after accounting for road rules (which provide a relatively concrete assessment of safety when
applicable) and controlling for differences in joint state, the data exhibits a high degree of variability
in safety labels. As a concrete example, we may consider weather which influences visibility (e.g.,
foggy conditions) and maneuverability (e.g., icy roads) which in turn should impact both the safety
measure and allowable action sets. In this case, we propose parameterizing the HJ reachability
computation on the interpretable axes of visibility and maneuverability.

**Responsibility.** Consider a road blockage which motivates a brief excursion into oncoming traffic.
Determining if this maneuver is allowable—let alone responsible—requires [D4] defining and in-
ferring responsibility from context and possibly noisy data. Existing methods have leveraged noisy
data to varying degrees in an effort to model and infer different notions of responsibility: as part of
an agent’s objective [16, 17] or control space (i.e., how much “effort” an agent puts into collision-
avoidance) [11, 18]. This naturally informs [D5] how to actionably incorporate responsibility into
a safety concept. By ensuring that the safety measure and allowable control sets are consistent, HJ
reachability serves as strong inductive bias when learning a cohesive combination of responsibility
representations; for example, learning a restriction on the control spaces $U^A$ and $U^B$, selecting ap-
propriate agent reactions ($\max / \min$), or automatically “shaping” the safety criterion $\ell(\cdot)$ via data.

### 4 Conclusion

For widespread adoption of AVs, stakeholders must converge on a common, rigorous safety concept.
We have proposed HJ reachability as a unifying mathematical framework for describing a family of
safety concepts, including those that already exist, to help stakeholders compare and contrast safety
concepts and converge to consensus. We hope that this document inspires the standardization of
novel safety-centric datasets and the use of HJ reachability as inductive bias when learning respon-
sibility subject to context, capturing two critical yet under-investigated aspects of safety.
References


