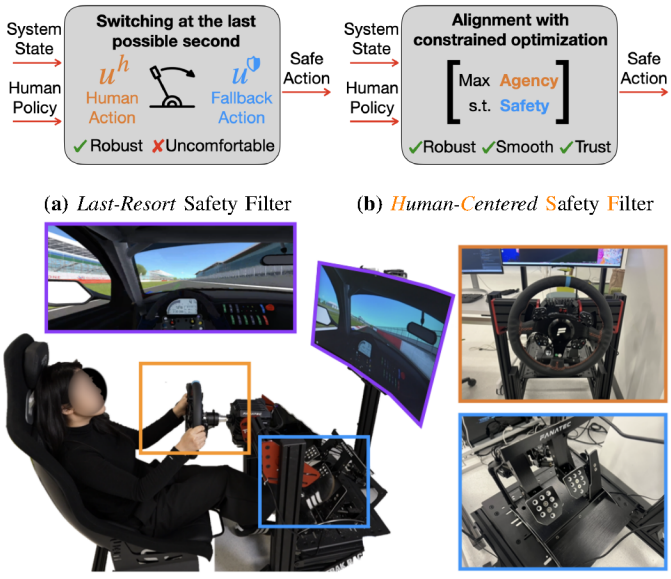
Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

Safety with Agency: Human-Centered Safety Filter with Application to AI-Assisted Motorsports

Donggeon David Oh[1,*], Justin Lidard[2,*], Haimin Hu[1], Himani Sinhmar[2], Elle Lazarski[1], Deepak Gopinath[3], Emily S. Sumner[3], Jonathan A. DeCastro[3], Guy Rosman[3], Naomi Ehrich Leonard[2], and Jaime Fernández Fisac[1]

***Abstract*—We propose a human-centered safety filter (HCSF) for shared autonomy that significantly enhances system safety without compromising human agency. Our HCSF is built on a neural safety value function, which we first learn scalably through black-box interactions and then use at deployment to enforce a novel state–action control barrier function (Q-CBF) safety constraint. Since this Q-CBF safety filter does not require any knowledge of the system dynamics for both synthesis and runtime safety monitoring and intervention, our method applies readily to complex, black-box shared autonomy systems. Notably, our HCSF's CBF-based interventions modify the human's actions minimally and smoothly, avoiding the abrupt, last-moment corrections delivered by many conventional safety filters. We validate our approach in a comprehensive in-person user study using Assetto Corsa—a high-fidelity car racing simulator with black-box dynamics—to assess robustness in "driving on the edge" scenarios. We compare both trajectory data and drivers' perceptions of our HCSF assistance against unassisted driving and a conventional safety filter. Experimental results show that 1) compared to having no assistance, our HCSF improves both safety and user satisfaction without compromising human agency or comfort, and 2) relative to a conventional safety filter, our proposed HCSF boosts human agency, comfort, and satisfaction while maintaining robustness.**



(a) *Last-Resort* Safety Filter **(b)** *Human-Centered* Safety Filter

(c) Shared Autonomy for High-Fidelity Car Racing

Fig. 1: Our proposed human-centered safety filter (HCSF) enables robust and smooth safety interventions for shared autonomy systems. (a) Last-resort safety filter (LRSF) switches to the best-effort fallback policy at the last possible moment. However, this switching can feel abrupt and uncomfortable for human operators. (b) Our HCSF instead intervenes smoothly while promoting human agency, thereby reducing automation surprise and enhancing user experience. (c) Users interact with a high-fidelity racing simulator via a steering wheel and set of pedals (throttle and brake).

I. Introduction

Recent developments in robot safety provide an exciting opportunity for enhancing human safety and performance in high-stakes situations. However, augmenting human decision-making with artificial intelligence (AI) in a trustworthy way remains an open problem. A human–AI team in a performance car racing [1–3] is a representative, yet challenging example, as it pushes safety to the limit. How should the AI co-pilot assist the human without dulling the driver's competitive edge? Should the AI discourage a human from attempting a risky overtaking maneuver on a sharp turn? When an AI system assists humans in such safety-critical and time-sensitive tasks, maintaining *human agency* is critical. We need to ensure human awareness of the AI system's current intent and operating mode, thus avoiding the notorious and sometimes fatal "automation surprise" [4, 5].

Safety filters [6, 7] have become an effective approach to ensure safety under an operational design domain (ODD), *i.e.*, a clearly defined set of operating conditions for robots to work properly and safely [8, 9]. Safety filters have been deployed on a wide range of autonomous systems, such as automated vehicles [10, 11], legged robots [12–14], and aerial navigation [15, 16]. Traditional model-based numerical approaches for safety filter synthesis [17] result in safety guarantees by design, but they are unable to scale up due to the "curse of dimensionality" [18]. Recent research focuses on neural approximation of safety filters [19–23] that can scale to tens [14, 24] and even hundreds of state variables [25]. While existing safety filters effectively maintain safety, their

[1]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08540, USA
[2]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08540, USA
[3]Toyota Research Institute, Cambridge, MA 02139, USA
This research has been supported in part by an NSF Graduate Research Fellowship. This work is partially supported by Toyota Research Institute (TRI). It, however, reflects solely the opinions and conclusions of its authors and not TRI or any other Toyota entity.
*D. D. Oh and J. Lidard contributed equally.

use in human–AI shared autonomy can result in abrupt, discontinuous interventions that disregard the human operator's intentions. This undermines the driver's sense of being in control, creating an uncomfortable and unenjoyable experience. Moreover, such unpredictable, non-transparent behavior can erode confidence and trust in the AI assistant, ultimately degrading team performance and causing the human driver to lose their strategic edge.

Contributions. To overcome these limitations, we propose a novel HCSF (Fig. 1) that advances the state of the art in learning-based safety filtering while actively promoting human agency in shared autonomy settings. In particular, we make three key contributions:

- We introduce, to the best of our knowledge, the first *fully model-free control barrier function (CBF) safety filter*. We learn a neural safety value function through interactions with a black-box system and, at deployment, enforce a safety constraint based on a novel *state–action control barrier function (Q-CBF)* without any knowledge of system dynamics (*e.g.*, control affine model). Both the synthesis and deployment of our Q-CBF are scalable to high-dimensional systems and do not require any knowledge of their dynamics.
- We build upon the learned Q-CBF and demonstrate our HCSF in Assetto Corsa (AC), a high-fidelity racing simulator with black-box dynamics, where the filter is pushed to the limit against all potential failure modes by real human drivers with diverse skill levels. To the best of our knowledge, this is the first time a safety filter has been synthesized, deployed, and evaluated in such a high-dimensional, dynamic shared autonomy setting involving human operators.
- We conduct an extensive in-person user study with 83 human participants and conclude, with statistical significance in both trajectory data and human driver responses, that our HCSF considerably improves safety and user satisfaction without compromising human agency or comfort relative to having no safety filter. Furthermore, when compared to a conventional safety filter, our HCSF offers significant gains in human agency, comfort, and overall satisfaction while maintaining at least the same level of robustness—if not exceeding it.

Overview. We organize this paper as follows. Section II reviews related works, while Section III introduces the problem formulation. In Section IV, we present our HCSF design, emphasizing its key properties and synthesis, and then discuss its practical implementation in Section V. Section VI provides our experimental results. Finally, Section VII addresses the limitations and outlines possible future directions, and Section VIII concludes the paper.

II. Related Work

Our work relates to, and builds on, recent advances in human-interactive safety filters and AI-assisted motorsports.

A. Human-Interactive Safety Filters

A safety filter is a supervisory control scheme that continuously monitors the operation of an autonomous system and intervenes, when necessary, by adjusting its planned actions to prevent potential catastrophic failures. Safety filters have been increasingly used in high-stakes autonomy applications, ranging from autonomous driving [11, 26–28], to aerial navigation [16, 20, 29–31], and to legged locomotion [12, 14, 23, 24]. Recent work by Hsu et al. [7] provides a unified analysis framework for various safety filters, including Hamilton–Jacobi (HJ) reachability [16–18], control barrier functions [20, 32, 33], model predictive control [34, 35], and Lyapunov methods [36]. In general, synthesis of safety filters can be computationally challenging, especially for systems with high-dimensional state space and complex dynamics. Deep learning has proven to effectively scale up the computation of safety controllers [14, 19, 23, 25, 37]. More recently, methods have been developed that treat these learned neural controllers as an untrusted fallback within a safety filter framework, and robust safety guarantees can be subsequently obtained through runtime verification algorithms such as convex optimization [38–40], forward reachable sets rollouts [22], and conformal prediction [41].

When robots are deployed around humans, ensuring safety is paramount to enable their trustworthy integration into people's everyday lives. However, enforcing safety becomes particularly challenging in human-interactive settings due to coupled motion, limited communication, and potentially conflicting objectives between robots and their human peers. Early attempts at safe human–robot interaction focus on achieving robust safety by safeguarding against worst-case human decisions [26, 42, 43], which may lead to overly conservative robot behaviors [44]. Recent research effort has been devoted to designing safety filters that adapt to human decision-making, in hope of improving the robot's task performance without compromising safety. One popular approach is filter-aware motion planning, which incorporates predictions of the safety filter's behaviors into the robot's task policy [26, 45]. This strategy allows the robot to avoid abrupt safety overrides by preempting future costly interventions triggered by unlikely human actions. Another line of research aims at reducing conservativeness by dynamically adjusting the safety filter's ODD according to the robot's evolving uncertainty about the human [25, 46, 47].

While existing human-interactive safety filters enable robots to interact safely and efficiently with *other* humans, similar formulations in human–robot *shared control* settings remain scarce. Recently, research efforts have focused on preserving human agency while enhancing safety in shared autonomy [48–50]. However, these approaches do not define a clear ODD and lack the principled safety analysis that a safety filter provides, often leading to elevated failure rates. Moreover, some methods rely on knowing the human operator's policy *a priori*, limiting their robustness when working with groups of human operators who have diverse intentions and skill levels

[48, 50].

In this work, we draw inspiration from filter-aware planning to design a safety filter that minimally modifies human actions. Our proposed HCSF preserves the principled safety analysis inherited from safety filter theory—in particular, from HJ reachability and control barrier functions—while avoiding any explicit model representation of human intentions.

B. AI-Enabled Motorsports

While modern AI systems surpass human intelligence in competitive sports [51, 52], their potential to *augment* human decision-making is underexplored. High-speed performance car racing presents a domain where safety and seamless collaboration are required to enable a competitive human–AI team—the AI co-pilot must assist the human without dulling the driver's competitive edge.

Wurman et al. [1] demonstrate for the first time that a well-trained neural policy can win a head-to-head competition against some of the world's best drivers in a car racing game. Follow-up works further improve the AI competitiveness via reasoning strategic interactions with data-driven modeling of opponent behaviors [53] and blending model-based dynamic game strategy with data-driven prior knowledge [54, 55]. Comparing to fully automated AI motorsports, human–AI collaborative car racing is an emerging, yet relatively underexplored research area. Gopinath et al. [3] propose a multi-task imitation learning approach that enables an automated coaching system that interacts with the student similar to a human teacher. DeCastro et al. [2] enhance the performance of human–AI teams in car racing by learning a policy that infers and aligns with human intents leveraging a world model. While AI agents in motorsports have shown promising performance, ensuring the safety of human drivers remains largely unaddressed in a principled manner. Chen et al. [56] present preliminary results on approximate learning-based safety analysis for autonomous racing, but their approach is limited to the single-car, fully automated setting.

This work presents an HCSF, a principled safety filter framework that actively promotes human agency, comfort, and satisfaction. We extensively evaluate our HCSF in a large-scale user study using AC, marking the first time a safety filter has been tested with both quantitative and qualitative measures of human–AI interaction in a high-fidelity, highly dynamic shared autonomy setting.

III. Preliminaries and Problem Formulation

We seek to ensure the safe operation of a robot with discrete-time nonlinear dynamics:

$$x_{t+1} = f(x_t, u_t), \tag{1}$$

where $x_t \in \mathcal{X} \subset \mathbb{R}^{n_x}$ and $u_t \in \mathcal{U} \subset \mathbb{R}^{n_u}$ denote the state and control input at time step $t \in \mathbb{N}$. The robot's control typically comes from a *task policy* $\pi^{\text{task}} : \mathcal{X} \to \mathcal{U}$. We define the *failure set* $\mathcal{F}$ with a Lipschitz continuous *safety margin function* $g : \mathcal{X} \to \mathbb{R}$:

$$\mathcal{F} := \{x \in \mathcal{X} \mid g(x) < 0\}. \tag{2}$$

States inside $\mathcal{F}$ are considered to have already failed in terms of safety. In the context of racing, states that correspond to the race car being outside the track boundaries or in contact with another vehicle should be inside $\mathcal{F}$. The control set $\mathcal{U}$ and failure set $\mathcal{F}$ are core components of the robot's ODD (Section V-C). The ODD may be understood as a social contract that bridges the robot operator, the public, and the policymakers—it provides a clear-cut set of conditions under which the robot is required to operate safely. To ensure safe robot operation under an ODD, we consider a supervisory control framework called safety filters.

A. Safety Filters

A *safety filter* [7] is an automated process that continuously monitors the system and intervenes, if deemed necessary, by modifying a candidate action given by the task policy π^{task} to prevent a potentially catastrophic safety failure *in the future*. Specifically, instead of directly applying the task action $u_t = \pi^{\text{task}}(x_t)$, the robot uses an action based on safety filtering:

$$u_t = \phi(x_t, \pi^{\text{task}}). \tag{3}$$

The specific function form of ϕ depends on the intervention type of a safety filter, which includes, *e.g.*, switching, transition, and optimization [7, Sec. 3]. The safety filter only prevents the use of a task action that would compromise future safety, allowing the robot to maintain safety without needing to modify its entire behavior. In this paper, we use HJ reachability analysis to synthesize safety filter ϕ [17, 18].

B. Hamilton-Jacobi Reachability Analysis

We aim to design a safety filter which, given ODD elements $\mathcal{F}$ and $\mathcal{U}$, keeps the robot within the *maximal safe set* $\mathcal{S}^* \subset \mathcal{F}^{\mathsf{c}} \subset \mathcal{X}$, where $\mathcal{F}^{\mathsf{c}} = \mathcal{X} - \mathcal{F}$. This set $\mathcal{S}^*$ consists of all states from which there *exists* a control policy that indefinitely prevents the robot from entering $\mathcal{F}$. In theory, $\mathcal{S}^*$ can be computed using Hamilton-Jacobi (HJ) reachability analysis, which employs level-set methods to recast the safety filter synthesis problem as an optimal control problem. Its solution follows from solving the dynamic programming *safety Bellman equation* [17]:

$$V(x) = \min\{g(x), \max_{u \in \mathcal{U}} V(f(x,u))\}, \tag{4}$$

which admits the *safety value function* $V : \mathcal{X} \to \mathbb{R}$ as its fixed-point solution. Given $V(\cdot)$, the maximal safe set $\mathcal{S}^*$ is then defined as:

$$\mathcal{S}^* := \{x \in \mathcal{X} \mid V(x) \geq 0\} \subset \mathcal{F}^{\mathsf{c}}. \tag{5}$$

For subsequent extension to our proposed HCSF, we adopt the $\mathcal{Q}$-function [57]—a notion widely used in reinforcement learning (RL)—to modify (4) into the *state–action safety Bellman equation*:

$$\mathcal{Q}(x,u) = \min\{g(x), \max_{u' \in \mathcal{U}} \mathcal{Q}(f(x,u), u')\}, \tag{6}$$

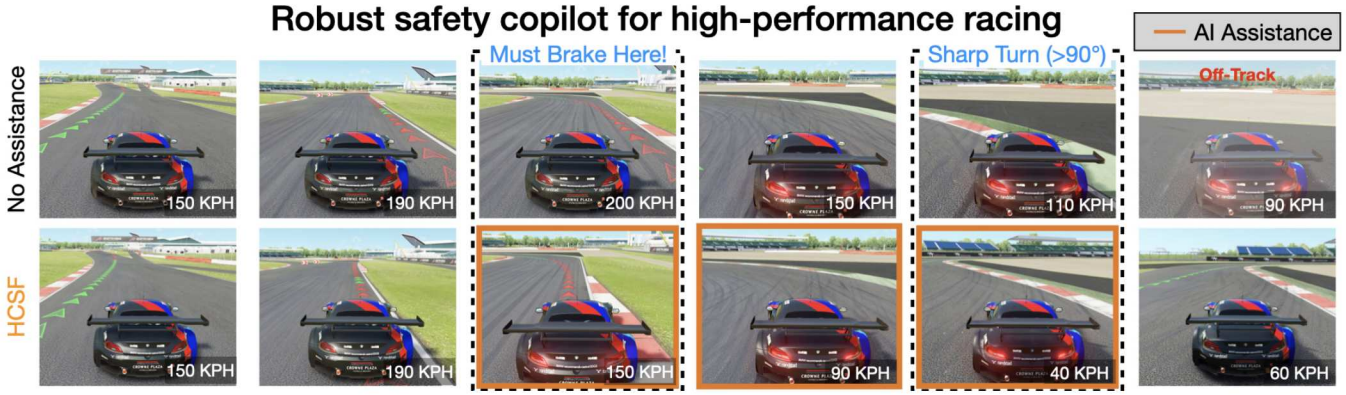


Fig. 2: Illustration of our HCSF intervention at a hairpin corner (*i.e.*, a sharp turn requiring rapid deceleration). Without safety filter assistance, inexperienced human drivers often miss the braking point, leading to understeering and the vehicle leaving the track. In contrast, our HCSF monitors the state and the human action to determine the braking point and provides necessary steering and braking interventions that keep the vehicle on the track. Braking assistance is visible through the rear lights.

which admits the *state–action safety value function* $\mathcal{Q} : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ as its fixed-point solution. This formulation remains equivalent to (4) in the sense that $V(x) = \max_{u \in \mathcal{U}} \mathcal{Q}(x, u)$.

We now introduce the *LRSF*, a value-based safety filter constructed upon the safety value functions $\mathcal{Q}(\cdot, \cdot)$ and $V(\cdot)$:

$$u(x) = \begin{cases} \pi^{\text{task}}(x), & \forall x \in \mathcal{X} \quad \text{s.t.} \quad V(x) > 0 \\ \pi^{\mathbf{0}}(x), & \text{otherwise,} \end{cases} \tag{7}$$

where the *safe fallback policy* is defined as $\pi^{\mathbf{0}}(x) := \arg\max_{u \in \mathcal{U}} \mathcal{Q}(x, u)$, and $\pi^{\text{task}}(\cdot)$ is any task-oriented policy that does not explicitly account for safety. We refer to (7) as a "last-resort" strategy since the filter does not intervene until $V(x) = 0$: the critical point at which the system is about to exit $\mathcal{S}^*$. Upon reaching the boundary of $\mathcal{S}^*$, LRSF fully overrides the control with $\pi^{\mathbf{0}}(\cdot)$ to enforce safety.

Prior work [16–19, 29] has established LRSF as a fundamental framework applicable for all HJ reachability analysis-based safety filters. This is due to its straightforward yet effective design for enforcing safety and its "least-restrictive" nature, allowing the task-oriented policy full freedom until the system reaches the boundary of the maximal safe set.

However, LRSF also has notable drawbacks. First, its *task-agnostic* fallback policy often leads to *discontinuous and jerky interventions* [7]. This issue can become more pronounced in a shared autonomy setting, where human operators might feel surprised and confused by abrupt overrides [4, 5]. Even though LRSF offers maximum freedom until the system reaches the boundary of $\mathcal{S}^*$, its interventions that do not take into account the human's input risk diminishing the operator's sense of control. Furthermore, the time and space complexities required to solve the Bellman equation (4) scale exponentially with the state-space dimension, rendering grid-based dynamic programming infeasible for real-world safe robot control.

Our proposed HCSF addresses these limitations by synthesizing an output that minimally deviates from the human operator's input, thereby enhancing both agency and smoothness while still enforcing safety (Section IV). In addition, in Section V, we leverage recent advances in safety RL [19, 23, 37] to approximate the $\mathcal{Q}$-function via RL, enabling the synthesis of a best-effort fallback policy $\pi^{\mathbf{0}}(\cdot)$ for high-dimensional systems.

C. Discrete-Time Control Barrier Functions

In this subsection, we introduce the definition and implementation of the discrete-time control barrier function (DCBF), another well-established approach for value-based safety filtering.

Definition 1 (Discrete-time CBF [13]). *A function $h : \mathcal{X} \to \mathbb{R}$ is a DCBF for system (1) if $\mathcal{S} = \{x \in \mathcal{X} \mid h(x) \geq 0\} \subset \mathcal{F}^{\text{c}}$ and $\exists \alpha \in (0, 1]$ that satisfies:*

$$\sup_{u \in \mathcal{U}} \Delta h(x, u) \geq -\alpha h(x), \quad \forall x \in \mathcal{X}, \tag{8}$$

where $\Delta h(x, u) := h(f(x, u)) - h(x)$.

Unlike LRSF, which imposes safety through hard overrides, a DCBF enables *smooth safety interventions* by solving an optimization problem that finds the safety-enforcing action closest to the task action $u^{\text{task}}(x)$:

$$u(x) = \underset{u \in \mathcal{U}}{\text{argmin}} \quad \left\| u^{\text{task}}(x) - u \right\|^2, \tag{9a}$$

$$\text{s.t.} \quad \Delta h(x, u) \geq -\alpha h(x), \tag{9b}$$

where (9b) is the *DCBF constraint*. The trade-off is that a DCBF no longer enforces safety within the maximal safe set $\mathcal{S}^*$ in general. Instead, it encodes safety with respect to a (smaller) safe set $\mathcal{S} = \{x \in \mathcal{X} \mid h(x) \geq 0\} \subseteq \mathcal{S}^*$.

An HJ safety value function $V(\cdot)$ is closely linked to a DCBF in the sense that, if $V(\cdot)$ is continuously differentiable, it automatically qualifies as a valid CBF for the maximal safe set $\mathcal{S}^*$ [7, Sec. 3.2]. This insight enables the use of $V(\cdot)$ in a smooth CBF safety filter rather than the LRSF alternative—an approach that underpins our proposed HCSF.

Remark 1. *Definition 1 could be relaxed such that (8) is required to hold for all $x_t \in \{x \in \mathcal{X} \mid h(x) \geq 0\}$. A control input $u_t \in \mathcal{U}$ that satisfies (9b) for any function $h(\cdot)$ meeting*

the relaxed DCBF definition still renders the 0-superlevel set of $h(\cdot)$ forward invariant. Such relaxation of (8) still guarantees safety, but loses the set attractiveness property for the 0-superlevel set of $h(\cdot)$ [58].

IV. Smooth Human-Centered Safety Filter for Shared Autonomy

In this section, we introduce a model-free, human-centered safety filter methodology that builds on HJ reachability analysis and DCBFs. We present our HCSF formulation and highlight its differences from existing safety filters.

Conventional CBF safety filter methods similar to (9) typically require knowledge of the system's dynamics [10, 13, 15, 31, 32, 58], even when using learned barrier functions [20, 33, 59–62]. This requirement arises for one or both of the following reasons: 1) either a full-order or a simplified dynamical model of the system is utilized to synthesize CBF candidates, and 2) knowledge of the dynamics (*e.g.*, control affine model) is leveraged at runtime to enforce the CBF safety constraint within an optimal control problem (OCP) (*e.g.*, quadratic program).

While some works explicitly aimed to build and deploy model-free CBF safety filters, they have so far fallen short of being *fully model-free*—relying on some combination of simplified models of the system dynamics [63–65], predefined low-level controllers [64, 65], and handcrafted fallback policies (*e.g.*, evading maneuvers) [66]. Additionally, recent efforts in learning a CBF for latent state representations have proven to be effective for partially observable systems, but they still require a control affine dynamical model [67]. Such reliance on knowledge of the system dynamics and the deployment environment can significantly limit the applicability of CBFs in complex, real-world scenarios where the dynamical model is often unknown and should be treated as a black-box. On the other hand, a model-free algorithm for learning a policy together with a barrier certificate was proposed recently [68], but it cannot be used to build a safety filter because the learned policy must be deployed at all times. Finally, we acknowledge a preprint reporting concurrent efforts toward a model-free state–action CBF safety filter [69]. However, it addresses a finite-horizon safety problem and enforces safety at runtime like a smooth least restrictive safety filter [70] rather than a CBF one, fundamentally differing from our work in both mathematical formulation and enforcement of safety.

To this end, we introduce, to the best of our knowledge, the first CBF safety filter that is *fully model-free*. We first show that the safety value function $V(\cdot)$ is itself a valid DCBF in the sense of Definition 1. Then, using the state–action safety value function $\mathcal{Q}(\cdot,\cdot)$ which could be learned scalably through black-box interactions with the system via model-free RL-based HJ reachability analysis (Section V-A), we propose a method of enforcing a novel Q-CBF safety constraint that does not require any information regarding the dynamics. While theoretical safety guarantees are contingent on the validity of the learned CBF (which may be established through statistical analysis [41] or model-based verification [20, 38]), this is not the focus of our research. Instead, we show our safety filter achieves an extremely high empirical safe rate and effectively preserves human agency.

We now present the Q-CBF formulation.

Proposition 1 (Q-CBF). *The safety value function $V(x) : \mathcal{X} \to \mathbb{R}$, which is a fixed-point solution of the safety Bellman equation (4), is a valid DCBF as defined in Definition 1 and Remark 1. The corresponding DCBF constraint is:*

$$\mathcal{Q}(x,u) \geq \gamma V(x), \tag{10}$$

where $\gamma \in [0,1)$. Following the notation from Definition 1, γ is equivalent to $1-\alpha$.

Proof: The proof is deferred to Appendix A. ∎

We emphasize the key difference between the original DCBF constraint (9b) and our formulation (10)—the requirement (or lack thereof) of the system dynamics. Given $\mathcal{Q}(\cdot,\cdot)$ that satisfies (6), (10) does not require the system dynamics for its evaluation. This enables its application to safety-critical systems with black-box dynamics, namely a high-fidelity car racing simulator. On the contrary, evaluating the original DCBF constraint (9b) does require *a priori* knowledge of the system dynamics, which prohibits it from being applied to systems with unknown dynamics. We also note that (10) constrains the system such that $V(\cdot)$ cannot decay below 0. In other words, it keeps the system within the maximal safe set $\mathcal{S}^*$, in contrast to many handcrafted DCBFs that often suffer from conservative safe sets.

We now leverage Proposition 1 to define our proposed HCSF, a safety filter tailored for shared autonomy settings (Fig. 2). Our HCSF solves an OCP at each timestep to find a safe action that minimally deviates from the human control $u^{\text{human}}(x)$ while satisfying the Q-CBF constraint.

Definition 2 (Human-Centered Safety Filter).

$$u(x) = \underset{u\in\mathcal{U}}{\operatorname{argmin}} \quad \left\|u^{\text{human}}(x) - u\right\|^2, \tag{11a}$$

$$\text{s.t.} \quad \mathcal{Q}(x,u) \geq \gamma V(x), \tag{11b}$$

where $\gamma \in [0,1)$ is a design parameter that dictates how quickly the safety value function is allowed to decrease over a single timestep.

The Q-CBF constraint ensures that the safety value function does not decay below the specified threshold $\gamma V(x)$ at each timestep. The recursive feasibility of HCSF is a direct consequence of $V(\cdot)$ being a valid DCBF, as stated in Proposition 1. This is formalized in the following proposition:

Proposition 2 (Recursive Feasibility of HCSF). *The optimization problem in Eq. (11) is recursively feasible for $\forall \gamma \in [0,1)$, given any initial state $x \in \mathcal{S}$.*

Proof: The proof is deferred to Appendix A. ∎

Together, Definition 2 and Proposition 2 establish that our HCSF *actively promotes human agency* by selecting an action that remains as close as possible to the human's intended input, while *ensuring the system never leaves the maximal safe set.*

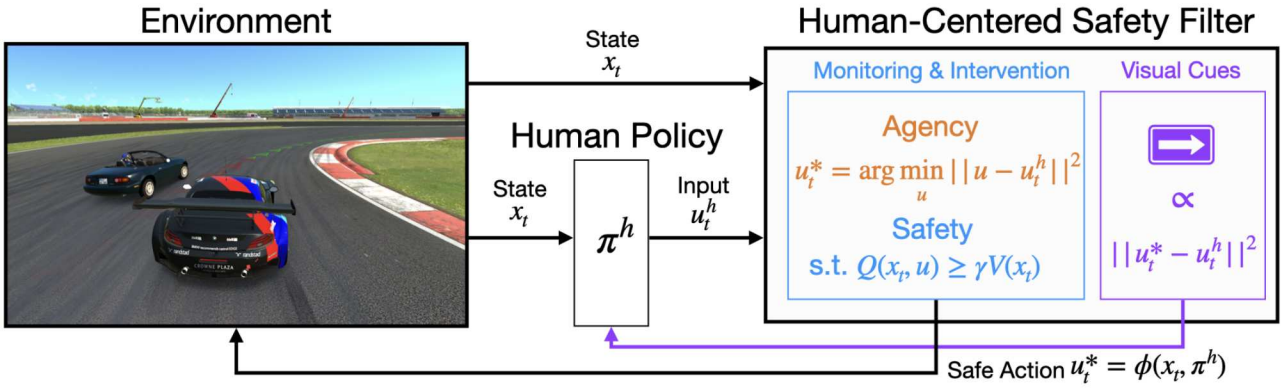


Fig. 3: A diagram describing the interaction between a human operator, our proposed HCSF, and AC game environment. Our HCSF utilizes a safety value function that we learn scalably through black-box interactions via model-free RL-based HJ reachability analysis, and at runtime leverages our novel Q-CBF constraint to enforce safety without any knowledge of system dynamics. Moreover, it intervenes minimally and smoothly to enhance human agency and comfort. Finally, our HCSF communicates the action modifications to the human driver via visual cues, facilitating transparent human–robot collaborartion.

Additional information regarding the practical implementation of our HCSF together with our choice of γ can be found in Appendix C. In the following section, we discuss a series of design choices that enable scalable and efficient implementation of the safety filters, along with details on their deployment in a high-fidelity car racing simulation.

V. Application of HCSF to High-Speed Racing

Our HCSF design in Section IV assumes the knowledge of the state–action safety value function $Q(\cdot, \cdot)$. However, directly solving (6) for high-dimensional systems is intractable—a manifestation of the notorious "curse of dimensionality." In this section, we leverage recent advances in safety RL to address this challenge and learn both the Q-function and the best-effort fallback policy in a high-fidelity racing environment. Specifically, we begin by describing the environment setup and then detail the observation, action, safety margin function, and episode termination conditions. Next, we outline our multi-phase training pipeline, which includes warmup and initialization phases that expedite learning by frequently exposing the agent to "dangerous" states. Finally, we discuss how we integrate safety filters with visual cues in an effort to enhance the AI's transparency.

A. Neural Synthesis of Safety Filters

Solving the safety Bellman equation (4) via dynamic programming is intractable for high-dimensional systems, as the computational and memory requirements grow exponentially with the dimensionality of the state space. Even state-of-the-art level-set methods can typically handle at most six continuous state dimensions, rendering them unsuitable for car racing applications.

Recent works [19, 22] in safety RL address this limitation by proposing a time-discounted variant of (4), which allows for scalable and effective approximation of the state–action safety value function $Q(\cdot, \cdot)$ and the best-effort fallback policy $u^{\mathbf{0}}(\cdot)$ via model-free RL algorithms, such as Soft Actor–Critic (SAC) [71].

During training, we accumulate a *replay buffer* $\mathcal{B}$ of transitions (x, u, g, x'), where $g := g(x)$. The critic (*i.e.*, state–action safety value function network) is then trained to predict the future discounted minimum margin by minimizing the Bellman residual, while the actor (*i.e.*, best-effort fallback policy network) is trained to maximize the safety value. Further details on the neural synthesis of safety filters are provided in Appendix C.

B. Human-Machine Interface

We use Assetto Corsa (AC), a high-fidelity racing simulator, together with a gym-compliant interface [72], to facilitate RL in a realistic driving environment. Specifically, we adopt the AC sim control interface (SCI), which integrates real-time hardware actuation (steering wheel and pedals) with trajectory data from the AC game engine and Python implementations of LRSF and HCSF. A first-person view is displayed on the monitor at 300 Hz. The SCI connects the hardware bus and the host computer via USB to run the safety filter loop (see Fig. 3), supporting a 30 Hz control rate—sufficient for a high-fidelity racing environment [1, 72].

The actuation platform consists of a Fanatec CSL DD QR2 wheel base, a Fanatec ClubSport Steering Wheel GT Alcantara V2, and Fanatec Clubsport Pedals V3. To enhance participants' immersion, we also include a Samsung S39C FHD 75Hz Curved Monitor and a Trak RS6 Racing Simulator rig.

C. Operational Design Domain (ODD)

We select the Silverstone Circuit (GP layout) for both training and deploying the safety filters, as its combination of fast straights and technical corners provides a challenging yet comprehensive proving ground. The ego vehicle is a BMW Z4 GT3, while Mazda MX-5 ND cars serve as opponent vehicles. Since the MX-5 ND is less powerful, it naturally encourages human drivers to attempt overtaking maneuvers. Although a single opponent is deployed during the user study, multiple opponents are used during training to increase on-track interactions and help the ego agent learn effective

collision-avoidance strategies. We use 50% opponent strength and 30% opponent aggression. The weather condition is set to "ideal", track conditions to "optimum", temperature to 26°C, and wind to 0 km/h. Additionally, traction control, stability control, and ABS are activated, while fuel consumption and tyre wear are turned off.

We learn the state–action safety value function $\mathcal{Q}(\cdot,\cdot)$ and the best-effort fallback policy $u^{\blacklozenge}(\cdot)$ based on the observation, action, margin function, and episode termination condition detailed below.

1) Observation: The system operates in a partially observable environment, where each observation is a 133-dimensional vector representing the ego agent's state and surroundings. This vector includes trajectory data (*e.g.*, speed, angular velocity, tire slip angles, distance to the reference path, and distances to track boundaries computed via ray-casting) from the last four timesteps, as well as the control inputs over the same four timesteps, in order to account for partial observability. The observation vector also contains the look-ahead curvature of the track and information about the nearest opponent, such as relative position, relative velocity, and braking status. A detailed breakdown of this 133-dimensional observation vector is provided in Appendix B.

2) Action: The normalized action space is defined as $\mathcal{U} = [-1, 1]^3$, with three continuous channels corresponding to steering, throttle, and brake. Gear changes are handled automatically via the gearbox feature provided by AC.

3) Margin Function: The margin function $g : \mathcal{X} \to \mathbb{R}$ is defined as the minimum between the signed distance to the track boundary and the signed distance to the nearest opponent, ensuring proximity-based safety constraints for both the environment and opponent vehicles. The corresponding failure set $\mathcal{F}$ is defined as in (2).

4) Episode Termination: If the margin function becomes negative or if the vehicle remains stationary for an elongated time period, the episode terminates and the vehicle is automatically reset to the closest point on the reference path. These episode termination conditions apply to both the neural synthesis of safety filters and the user study.

D. Warmup and Initialization

In AC, resetting the vehicle places it stationary on the closest point of the reference path, making it difficult to gather training data for near-failure scenarios where safety filters are most critical. To address this, we use a two-phase pipeline (warmup and initialization) that accelerates the vehicle to higher speeds under a performance-oriented policy (warmup), then systematically pushes it into more challenging or hazardous states (initialization). By deliberately inducing these "dangerous" situations (including adversarial and random maneuvers near the boundary of the safe set), our pipeline ensures the agent encounters a wide range of conditions where the safety filters must intervene effectively. This design both reduces wall-clock training time by avoiding trivial low-speed states and promotes robust learning, as the filter gains experience in precisely the situations where safety intervention is needed most. Full details on the warmup and initialization phases can be found in Appendix C.

Fig. 4: Our HCSF displays two types of visual cues: horizontal arrows that reflect the modifications made to the steering input, and vertical arrows that indicate the corrections made to the throttle/braking inputs. The length of each arrow is proportional to the magnitude of modification made to the corresponding input channel.

E. Training Details

We train the policy and value networks on a single RTX 4090–equipped machine with an AMD Ryzen 9 7950X 16-core processor. A replay buffer of size 20 million is used, and the actor and critic networks are each updated once per environment step. Both the policy and value networks are three-layer MLPs with 256 neurons per hidden layer, trained with a batch size of 128. The networks are trained for over three weeks (12.8 million environment steps) using the Adam optimizer. We use the same neural approximation of $\mathcal{Q}(\cdot,\cdot)$ for all safety filters. Further details on training hyperparameters are provided in Appendix C.

F. Visual Cues

In our framework, low-bandwidth visual cues foster transparency and collaboration between the human driver and a safety filter. Whenever an intervention occurs, vertical and horizontal arrows on the screen show both the direction and magnitude of the AI's corrections relative to the driver's original input. Specifically, the arrow's orientation indicates whether the AI is steering more to the left or right compared to the human, or braking more or less compared to the human, while the arrow's length is proportional to the magnitude of that difference. By mapping each cue to a distinct control channel, we avoid unnecessary information loss. We also omit audio cues, which could add cognitive load or distraction in high-speed scenarios. As illustrated in Fig. 4, this setup allows the driver to immediately recognize when and how the system intervenes.

During our user study, all participants were shown a color-coded reference path, which is a series of green and red arrowheads on the track (Fig. 2). Green arrowheads indicate acceleration, and red arrowheads advise deceleration. This visual aid allows drivers to navigate effectively without prior expertise in sim racing.

VI. Experimental Results

In this section, we present experimental results that provide evidence for our hypotheses:

- **H1**: Our HCSF improves **safety** and user **satisfaction** without compromising human **agency** and **comfort**, compared to having no filter.
- **H2**: Our HCSF improves human **agency**, **comfort**, and **satisfaction** without compromising **robustness**, compared to LRSF.

To test these hypotheses, we conducted a large-scale user study in the AC simulation environment described in Section V-C. This study involved 83 participants with diverse driving backgrounds, marking the first time the interaction between human operators and safety filters has been systematically investigated.

A. Baselines

We compare our proposed HCSF against two baselines—*LRSF* and *unassisted driving*—by examining both quantitative trajectory data and qualitative user experience.

For LRSF, we follow the formulation in (7). This filter only intervenes when the system reaches the boundary of the safe set and does not consider the human operator's input during interventions. We anticipate that such an abrupt approach that disregards human intent may confuse drivers and undermine comfort, thus increasing the risk of automation surprise. In contrast, our HCSF solves the optimization problem (11) to minimally modify the human action while still satisfying the Q-CBF constraint, which we hypothesize will enhance human agency, comfort, and overall satisfaction compared to LRSF. To ensure a fair comparison between our HCSF and LRSF, we employ the same neural approximation of the state–action safety value function $\mathcal{Q}(\cdot,\cdot)$ for both value-based safety filters. This ensures that *both filters rely on the same safety monitor, while their interventions may differ* (7), (11). In addition, *each filter's visual cues follow the same proportionality rule*, maintaining consistency in how interventions are conveyed to human drivers.

We also include a control group that receives no safety filter assistance. This allows us to capture any unassisted learning effect—where participants may improve simply through practice—and to gauge the placebo effect of believing one might be assisted by AI, even when no assistance is provided. To control for this placebo effect, all participants are informed that they may receive AI assistance, regardless of their actual assignment. Given that our HCSF should substantially reduce accidents, we hypothesize it will achieve superior safety and user satisfaction compared to the unassisted group. Moreover, thanks to its smooth, human-centered interventions, we expect our HCSF to preserve human agency and smoothness relative to having no safety filter at all.

B. Metrics

We evaluate our core hypotheses using four core metrics: *robustness*, human *agency*, *comfort*, and overall *satisfaction*. In addition, we assess four *filter-specific metrics*—trustworthiness, predictability, interpretability, and competence—to gain further insight into human–safety filter interactions, even though these filter-specific measures are not directly tied to our hypotheses.

1) Robustness: We evaluate the robustness of safety filter–assisted decision making for human drivers using three forms of quantitative trajectory data: out-of-track incidents (per minute), collisions (per minute during close-proximity interaction), and failures (per minute).

Since out-of-track incidents and collisions are two different modes of failure, we normalize them separately. Out-of-track incidents can occur at any moment in a driving session (*e.g.*, a driver might instantly veer off track by sharply flicking the steering wheel), so we divide the total count by the session length. By contrast, collisions can only happen when the ego vehicle is close to an opponent, so we normalize the collision count by the total time spent within a specified distance threshold.

We also gather qualitative data on how robust participants perceive the interaction to be. Even if the quantitative trajectory data indicate strong robustness, drivers may not necessarily feel confident. For example, while many modern vehicles feature lane keeping assist (LKA) systems that effectively reduce lane departures, they can cause a vehicle to bounce between adjacent lane markings. Some drivers lose confidence because of this "bouncy" behavior, leading them to disable the feature despite its technical effectiveness [73]. Therefore, we ask participants whether they feel confident in their ability to drive safely throughout each session.

2) Agency: Prior work in cognitive psychology [74–76] and human–robot interaction (HRI) [77–79] interprets human agency as the correspondence between intended and actual actions. This suggests that agency, in the context of safety filtering, is better framed as a *game of degree* (*i.e.*, how much the system intervenes), rather than a *game of kind* (*i.e.*, whether it intervenes at all). Consistent with this interpretation, we broadly define human agency as the *degree* of control that the driver has over the vehicle. To quantify agency, we define the input modification (I.M.) measure as the ℓ_2-norm of the difference between the human operator's raw input and the final input applied after safety filtering:

$$\text{I.M.}(t) = ||u_t^{\text{human}} - \phi(x_t, u_t^{\text{human}})||_2, \tag{12}$$

where ϕ can be any of the three safety filters in {HCSF, LRSF, none}.

In addition to this quantitative trajectory data, we also gather qualitative data on human agency by asking participants how much control they felt they had over the vehicle.

3) Comfort: We measure human comfort using two key forms of quantitative trajectory data: the jerk magnitude and the magnitude of the first-order control input difference.

First, we define the jerk magnitude as:

$$\text{jerk}(t) = ||\dddot{p}_t||_2, \tag{13}$$

where p_t is the vehicle's position in three-dimensional Euclidean space. Large jerk values can lead to discomfort or

motion sickness, whereas smaller jerk values typically indicate a smoother, more comfortable ride [80, 81].

Next, the squared magnitude of the first-order control input difference (I.D.) is given by:

$$\text{I.D.}(t) = ||\phi(x_t, u_t^{\text{human}}) - \phi(x_{t-1}, u_{t-1}^{\text{human}})||_2^2, \tag{14}$$

where smaller I.D. values are associated with better passenger comfort [82]. In this study, we use I.D. to compare the smoothness of input trajectories obtained from our human-centered optimization problem (11) with those produced by the best-effort fallback policy (7).

Finally, we also collect qualitative feedback by asking participants how "smooth" they perceived the driving experience to be. We use the term "smooth" instead of "comfortable" for two reasons: 1) since participants receive no actual motion feedback, it is difficult for them to assess ride comfort, and 2) "comfortable" can be mistaken for "confident," given their similarity in everyday usage, which could introduce unwanted ambiguity in the responses.

4) Satisfaction: Because overall satisfaction cannot be measured through quantitative trajectory data, we rely entirely on qualitative data obtained from participants regarding their satisfaction with the overall driving experience.

5) Filter-Specific Metrics: For additional insights into how human drivers interact with the safety filters, we collect qualitative data on the trustworthiness, predictability, interpretability, and competence of the safety filter assistance. We refer to these four items as "AI-specific metrics" because they explicitly address the interventions made by the safety filter, whereas our four core metrics focus on the broader driving experience during a session. For instance, it makes sense to evaluate the smoothness of a session even if no safety filter is deployed, but asking about the trustworthiness of an intervention is only applicable when a filter is actively involved.

C. User Study Design

In this subsection, we describe the design of our user study, including details on participant recruitment, group assignments, and experiment procedure.

1) Participant Recruitment: We reached out to potential participants both online and offline. Specifically, we used university-wide mailing lists, social media advertisements, and printed posters to engage faculty, staff, and students at our institution, aiming to minimize biases related to age, gender, or academic major and to ensure a wide range of perspectives. No monetary or academic compensation was provided to participants.

2) Group Assignments: We consider prior driving and video gaming experience to be the most influential factors for our study's results and explicitly controlled for them. We asked participants to report their driving experience level on a five-point scale, following these criteria:

1) I have no experience in either real or simulated driving.
2) I have driving experience but no racing experience.
3) I have some experience in racing, but only in simulation.
4) I have extensive racing experience, but only in simulation.
5) I have real car racing experience.

We assigned participants to each group so that their average initial skill levels remained comparable, ultimately producing group sizes of 25–29 participants. Table I presents the mean and standard deviation of each group's reported driving experience. A one-way analysis of variance (ANOVA) followed by Tukey's honestly significant difference (HSD) indicates that p-values for comparisons between any two groups exceed 0.80, suggesting no statistically significant differences in initial skill levels across groups.

	HCSF	LRSF	None
Number of Participants	29	29	25
Average Initial Skill level	2.17 ± 0.54	2.21 ± 0.86	2.28 ± 0.68

TABLE I: Number of participants in and the average initial skill level of each group.

3) Experiment Procedure: In this study, each participant completes three separate driving sessions. During the first session, participants drive for five minutes without any safety filter assistance. This initial session establishes a baseline for each participant's initial skill level, complementing the group assignment procedure, which also ensures that average initial skill levels remain similar across all groups. All participants are explicitly informed that no assistance is provided in this session. In the second session, participants drive for ten minutes under the assistance of the safety filter corresponding to their assigned group, although they are only told that they may receive AI assistance (with no specification of its type). Finally, as in the first session, participants drive for five minutes without assistance in the third session. This third session is intended to reveal potential over-reliance on safety filters; if participants fully trust the robustness of the filter during the second session, they might rely heavily on it, resulting in marginal or no improvement in their own driving skills. By comparing trajectory data across all three sessions, we can analyze the possibility of over-reliance.

Immediately after each session, participants answer questions regarding the four core metrics: robustness, agency, comfort, and satisfaction. The filter-specific metrics (trustworthiness, predictability, interpretability, and competence) are only queried after the second session, when participants may actually experience safety filter interventions. Each metric is measured using two statements: an affirmative form and a negated form. This "reverse-coded" approach helps detect inattentive or biased responses and ensures that the underlying construct is captured from different angles, thus enhancing the reliability of the qualitative measures. We verify that each pair of affirmative and negated statements measures the same construct by conducting a Cronbach's alpha test, the results of which are reported in Table VII in the Appendix. All items use a five-point Likert scale. To interpret a participant's overall response to a given metric, we average the rating from the affirmative statement with $6 -$ (the rating from the negated statement).

Driving Session	Time	Assistance
Session 1	5 minutes	None
Session 2	10 minutes	LRSF, HCSF, or None
Session 3	5 minutes	None

TABLE II: Duration of and type of assistance provided in each session.

D. Results

In this subsection, we present the results of our user study. We validate our hypotheses on a metric-by-metric basis, using both qualitative and quantitative data for the four core metrics: robustness, agency, comfort, and satisfaction. We also analyze qualitative responses for filter-specific metrics to gain further insight into how participants interact with each safety filter.

When analyzing quantitative or qualitative data that follow a two-factor (session and group) repeated-measures design—i.e., data collected across multiple sessions with an interest in comparing different groups—we employ a Mixed ANOVA model to account for both within-subject (session) and between-subject (group) factors. Whenever a significant interaction between session and group emerges in the Mixed ANOVA test, we proceed to the simple main effects (SME) analysis to check how the groups differ separately in session 1 and session 2. Then, if the SME analysis results in statistical significance, we perform Tukey's HSD test to pinpoint precisely how each group's distribution of data diverges from others within a session, thereby clarifying the role of the safety filters and how they influence each metric.

For data that do not meet this two-factor repeated-measures structure (i.e., data not measured over multiple sessions), we use a one-way ANOVA that considers only the between-subject (group) factor. If the ANOVA result indicates statistical significance, we then apply Tukey's HSD test to compare the distributions pairwise for each group combination.

For all numerical data depicted in bar or box plots, we illustrate statistical significance using asterisks to compare pairs of groups. Formally, under the null hypothesis H_0, we assume the data from any two groups come from the same underlying distribution. If a statistical test (*e.g.*, Mixed ANOVA followed by SME and Tukey's HSD) indicates that $p < 0.05$, we reject H_0 at the 5% level and mark the pair with one asterisk ($*$). Likewise, we use two asterisks ($**$) when $p < 0.01$, and three asterisks ($***$) when $p < 0.001$. Thus, more asterisks correspond to stronger evidence against H_0 and hence a more statistically significant difference between the groups' data.

1) Robustness: As shown in Fig. 5, our HCSF maintained near-zero failures across the study, demonstrating strong robustness under diverse human actions. It significantly reduced out-of-track incidents per minute and overall failures per minute compared to the unassisted group. While our HCSF also reduced out-of-track incidents and collisions relative to LRSF, these differences were not statistically significant.

A similar trend emerges in the qualitative results. As shown in Fig. 6, participants assisted by our HCSF in session 2 felt significantly more confident in the safety filter's robustness than those in the unassisted group. In addition, the difference in each participant's confidence score between sessions 1 and 2 was significantly greater for our HCSF than for unassisted driving. Although participants in the HCSF group also reported higher confidence than those in the LRSF group, this difference was not statistically significant.

Hence, we conclude:

- Our HCSF dramatically improves safety in decision-making compared to having no filter.
- Our HCSF is at least as robust as LRSF, if not more.

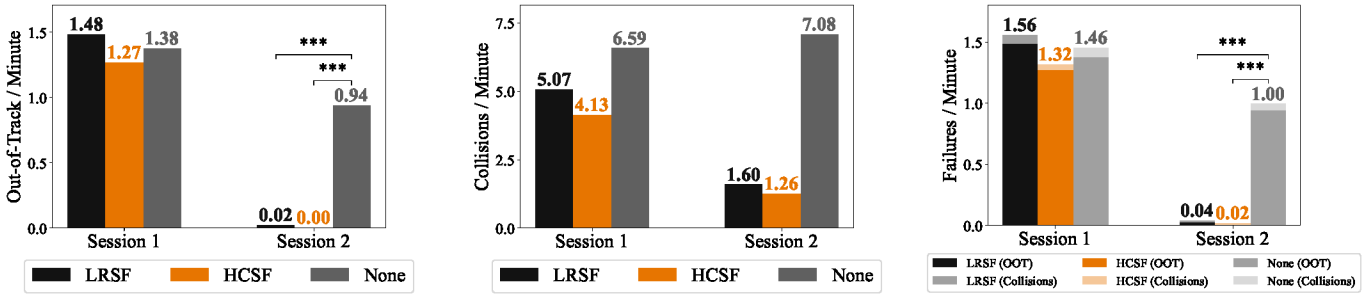
Although our HCSF did reduce collisions relative to LRSF and unassisted driving, there was no statistically significant difference. We conjecture that this may be due to two main causes. First, during the training of the neural approximation of the state–action safety value function and the best-effort fallback policy, we did not treat the opponent as an adversarial agent; instead, we considered the opponent as a part of the environment. We did so because treating the opponent as adversarial would prohibit side-by-side racing—making overtaking impossible—since the opponent could easily induce a collision by side-swiping the ego vehicle. Additionally, AC does not allow users to dictate actions for opponent vehicles, thus preventing the setup of an adversarial multi-agent RL. In other words, in order for a car race to occur, we cannot guarantee absolute robustness against collisions. Second, participants frequently attempted overtakes in unorthodox parts of the track, where real-world racing discourages such maneuvers. Although they might successfully overtake without colliding, they often carried excessive speed and ended up driving off-track. Consequently, the number of collisions may have been underreported in session 1 and for the unassisted group during session 2.

2) Agency: Fig. 7 reports the *distribution of I.M. among all participants in each group across all timesteps*—including those when no safety filter intervention occurred—for the HCSF and LRSF groups, thus jointly capturing intervention frequency and magnitude. Notably, our HCSF reduces the frequency of interventions with I.M.$>$1.0 relative to LRSF, so the filtered actions remain closer to the human operator's intended actions. While our HCSF intervenes more frequently than LRSF (30.3% vs. 19.7% of the driving session duration), its interventions are much smaller in magnitude (0.184 vs. 0.305 in average I.M.). This result shows that our HCSF significantly improves human agency by reducing the input modification magnitude over LRSF. By contrast, LRSF does not take into account the human action, which can lead to unnecessarily large, abrupt corrections that undermine the driver's sense of control. This tendency is even more pronounced in the empirical cumulative distribution function (ECDF) of I.M. in Fig. 8a, and Fig. 8b shows that the HCSF group has a significantly lower average I.M. than the LRSF group.

Qualitative measures also strongly support the improved human agency of our HCSF over LRSF. As shown in Fig. 9, participants assisted by our HCSF during session 2 reported a significantly stronger sense of being in control compared

Metric	Related Questions
Robustness	**affirmative**: I felt confident that I could drive safely throughout the race. **negated**: I felt nervous about whether I would be able to finish the race without accidents.
Agency	**affirmative**: I felt I was in control of the vehicle throughout the race. **negated**: There were times when the vehicle wasn't doing what I wanted.
Comfort	**affirmative**: The driving experience felt smooth overall. **negated**: There were times when the drive felt jerky.
Satisfaction	**affirmative**: Overall, I am happy with how the race went. **negated**: This race did not go as well as I thought it could have.
Trustworthiness	**affirmative**: I trusted the AI Assistant to keep me safe throughout the race. **negated**: I felt uneasy about whether the AI Assistant would get me into an accident.
Predictability	**affirmative**: The AI Assistant's actions were predictable. **negated**: The AI Assistant's actions surprised me at times.
Interpretability	**affirmative**: The AI Assistant's actions made sense to me. **negated**: Sometimes, I couldn't figure out why the AI Assistant was taking actions.
Competence	**affirmative**: The AI Assistant seemed to have a good grasp of the situation. **negated**: Sometimes, the AI Assistant didn't seem to know what it was doing.

TABLE III: 4 core metrics and 4 filter-specific metrics together with their corresponding affirmative/negated questions.



(a) Out-of-track incidents per minute. (b) Collisions per minute ($< 3m$ distance). (c) Total number of failures per minute.

Fig. 5: Our HCSF achieved near-zero failures throughout the user study, demonstrating significant enhancement in safety compared to unassisted human driving. Although our HCSF outperformed LRSF in both failure modes, the differences were not statistically significant. Statistical significance is marked with asterisks, where more asterisks indicate larger significance.

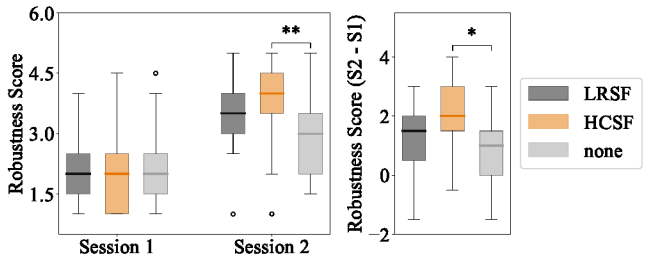


Fig. 6: Qualitative measures of robustness across all three groups in both sessions. Participants in the HCSF group reported significantly higher confidence in session 2 compared to the unassisted group. Central marks, bottom, and top edges of the boxes indicate the median, 25th, and 75th percentiles, respectively. The maximum whisker length is set to 1.5 times the standard deviation, which gives 99.7 percent coverage for normally distributed data. Statistical significance is marked with asterisks, where more asterisks indicate larger significance.

to those assisted by LRSF. This difference is even more pronounced if we consider the change in each participant's agency score between sessions 1 and 2. Meanwhile, the unassisted group recorded the highest average agency score overall—unsurprising since they had full control of the vehicle in both sessions—but its difference from the HCSF group was not statistically significant in both session 2 scores or the improvement across sessions. Finally, the LRSF group exhibited a drop in agency from session 1 to session 2, underscoring how LRSF undermines the human operator's sense of control. By contrast, our HCSF manages to preserve agency at a level comparable to having no filter at all.

Therefore, we conclude:

- Our HCSF does not compromise human agency compared to having no filter.
- Our HCSF significantly enhances human agency compared to LRSF.

3) Comfort: Although our HCSF does not explicitly address comfort in the optimization problem (11), we hypothesize that it will still provide a smoother ride compared to LRSF. We expect our HCSF to minimize deviations from the human input (11a), effectively "anchoring" the filtered action around the human operator's commands. Because humans are physically limited in how quickly they can turn the wheel or press the pedals, their inputs naturally form smooth

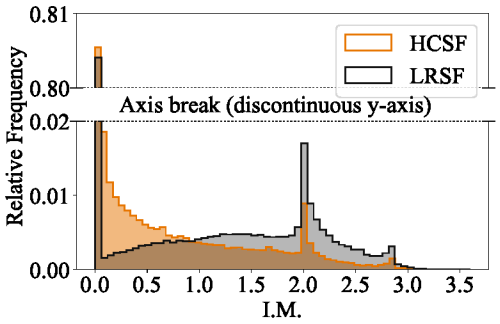
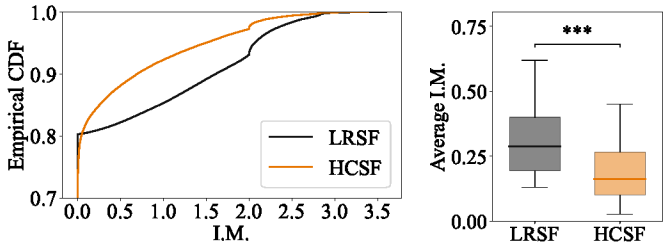


Fig. 7: Histogram of I.M. for the HCSF and LRSF groups over all timesteps, including those when no safety filter intervention occurred. Compared to LRSF, which often produces large input modifications that undermine human agency, our HCSF reduces the frequency of such large modifications by offering human-centered "nudges" of smaller magnitude.



(a) ECDF plot of I.M. **(b)** Box plot of average I.M.

Fig. 8: (a) For each group, the presented ECDF aggregates the I.M. values from all participants across every timestep, including those when no safety filter intervention occurred. (b) The box plot summarizes the distribution of each participant's average I.M. within each group. Our HCSF yielded a significantly smaller I.M. compared to LRSF, suggesting better retention of human agency.

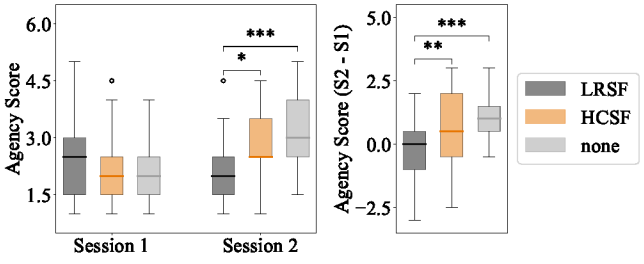


Fig. 9: Qualitative measures of human agency across all three groups in both sessions. Participants in the HCSF group reported a significantly stronger sense of being in control compared to those in the LRSF group, while no significant difference emerged between the HCSF and unassisted groups.

and continuous control trajectories. As a result, the resulting control actions under our HCSF assistance are anticipated to inherit some of that smoothness, thus translating into non-jerky movement. In contrast, LRSF, which disregards the human input, can cause larger discontinuities in the control signals when switching abruptly between human control and the best-effort fallback policy. Moreover, the best-effort fallback policy itself is a neural network trained with no explicit smoothness reward, further contributing to potentially jerky control.

The ECDF plot in Fig. 10a illustrates the I.D. distribution among all participants in each group across all timesteps where a safety filter intervened (and hence no data for the unassisted group). We observe our HCSF producing a denser distribution at smaller I.D. values. In other words, our HCSF tends to yield smoother inputs than LRSF. The box plot in Fig. 10b reinforces this observation by showing a significant difference in average I.D. distribution between the LRSF and unassisted groups. Moreover, although the unassisted group exhibits the smallest average I.D., its difference from the HCSF group is not statistically significant.

A similar pattern appears in Fig. 11a, which shows the jerk distribution among all participants in each group across all timesteps where a safety filter intervened (and hence no data for the unassisted group). Here, our HCSF produces a denser distribution at smaller jerk compared to LRSF, indicating smoother, more comfortable driving. The box plot in Fig. 11b further confirms that our HCSF significantly reduces average jerk relative to LRSF, while its difference from unassisted driving is not statistically significant.

Finally, Fig. 12 illustrates the participants' sense of smoothness. Those assigned with LRSF reported a decline in their sense of smoothness from session 1 to session 2, indicating discomfort arising from abrupt and discontinuous LRSF interventions. In contrast, participants assisted by our HCSF reported a significantly higher smoothness score in session 2 compared to those assisted by LRSF, and this difference becomes even more pronounced when examining the change in each participant's score between sessions 1 and 2. Similar to the agency results, the unassisted group exhibited the highest average smoothness score overall, which is unsurprising given the absence of interventions of any sort. Nonetheless, our focus here is on how well our HCSF preserves a smooth driving experience; the gap between the HCSF and unassisted groups was not statistically significant in either the session 2 scores or in the improvement across sessions.

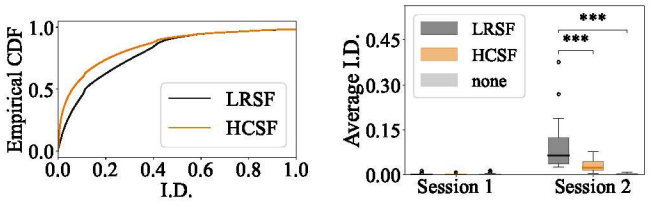
Therefore, we conclude:

- Our HCSF does not compromise comfort compared to having no filter.
- Our HCSF significantly improves comfort compared to LRSF.

4) Satisfaction: Fig. 13 shows that during session 2, participants who received our HCSF assistance reported significantly higher overall satisfaction scores compared to those who received LRSF assistance or no assistance. This difference in user satisfaction between the HCSF and LRSF groups is also evident when examining each participant's change in scores from session 1 to session 2.

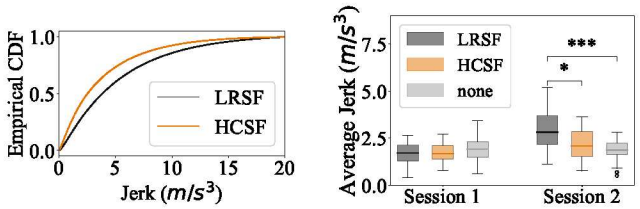
Therefore, we conclude:

- Our HCSF significantly improves user satisfaction compared to having no filter.
- Our HCSF significantly improves user satisfaction compared to LRSF.



(a) ECDF plot of I.D. (b) Box plot of average I.D.

Fig. 10: Our HCSF yielded a significantly smaller I.D. compared to LRSF, indicating smoother input trajectories, while its difference from unassisted driving was not statistically significant. Central marks, bottom, and top edges of the boxes indicate the median, 25th, and 75th percentiles, respectively. The maximum whisker length is set to 1.5 times the standard deviation, which gives 99.7 percent coverage for normally distributed data. Statistical significance is marked with asterisks, where more asterisks indicate larger significance.



(a) ECDF plot of jerk. (b) Box plot of average jerk.

Fig. 11: Our HCSF produced significantly smaller jerk compared to LRSF, suggesting better ride comfort, while its difference from unassisted driving was not statistically significant.

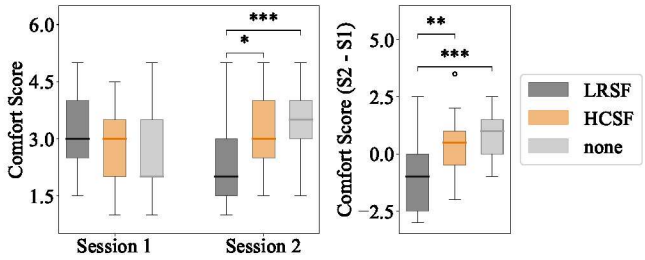


Fig. 12: Qualitative measures of comfort across all three groups in both sessions. Participants in the HCSF group reported a significantly better sense of smoothness than those in the LRSF group, while the difference between the HCSF and unassisted groups was not statistically significant.

5) Filter-Specific Metrics: We further analyze the interplay between human drivers and the safety filter using four filter-specific metrics—trustworthiness, predictability, interpretability, and competence. The participant responses, shown in Fig. 14, reveal that our HCSF exhibits significantly higher trustworthiness and competence compared to unassisted driving. Meanwhile, LRSF falls between our HCSF and unassisted driving without a statistically significant difference. Additionally, participants rated LRSF as significantly more unpredictable compared to unassisted driving.

Although the quantitative robustness measures in Fig. 5 show no significant difference between our HCSF and LRSF,

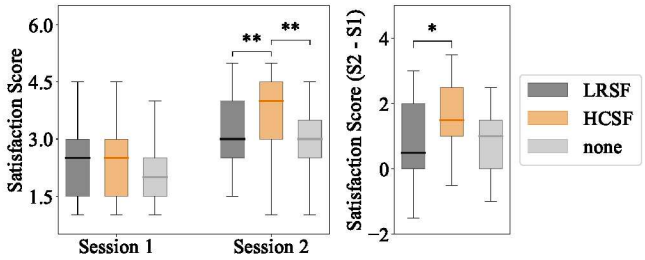


Fig. 13: Qualitative measures of overall satisfaction across all three groups in both sessions. Participants in the HCSF group reported significantly higher satisfaction during session 2 compared to both the LRSF and unassisted groups.

we conjecture that participants perceived LRSF as less trustworthy and competent due to its disregard for human input, which results in discontinuous, jerky, and unpredictable interventions. Unable to anticipate when or how LRSF would intervene, participants were less inclined to trust the system. In contrast, because our HCSF minimizes deviations from the human actions while still preserving robustness, it may have felt more trustworthy and competent to the users.

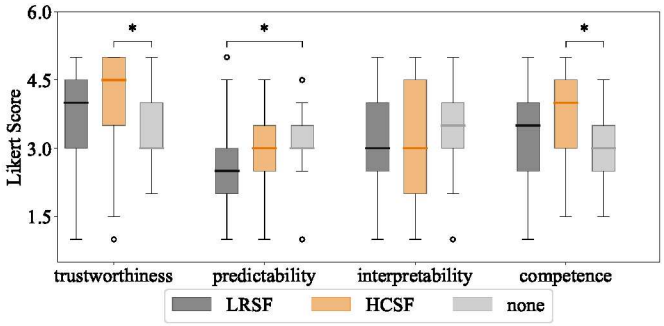


Fig. 14: Qualitative measures of filter-specific metrics—trustworthiness, predictability, interpretability, and competence—across all three groups in session 2. Central marks, bottom, and top edges of the boxes indicate the median, 25th, and 75th percentiles, respectively. The maximum whisker length is set to 1.5 times the standard deviation, which gives 99.7 percent coverage for normally distributed data. Statistical significance is marked with asterisks, where more asterisks indicate larger significance.

VII. Limitations and Future Work

Our proposed HCSF offers significant improvements in human agency, comfort, and satisfaction compared to conventional safety filters; however, it does have several limitations. First, our HCSF does not directly optimize for racing performance (*e.g.*, lap times), as its sole objective is to minimize deviation from the human input while satisfying the Q-CBF constraint. Although our HCSF offers a slight performance gain over LRSF and unassisted driving, the difference in lap times was not statistically significant (Appendix D). Second, prolonged exposure to our HCSF may result in drivers becoming *over-reliant* on AI assistance, potentially hindering the development of unassisted driving skills (Appendix D).

Third, the inherently reactive design of our HCSF addresses only immediate safety threats, leaving drivers unprepared for complex race dynamics that require proactive anticipation and adaptation. Finally, while visual cues can help reduce confusion during safety interventions, they may also induce unintended behavioral changes that merit further investigation.

To address these limitations, future work could add a secondary intervention layer that activates prior to safety filter interventions and provides proactive, multi-modal cues instead of removing the human operator's control authority. We expect this new layer to mitigate the potential over-reliance on safety filters because it does not intervene at the physical control channels shared between a human operator and a safety filter; instead, it "coaches" the operator to take appropriate actions before a safety intervention is needed. Moreover, since we can rely on a safety filter to maintain system safety, the new layer can be designed to foster performance-oriented operation. Building this additional layer will require further investigation into how humans respond to different modalities and content of cues. Overall, these advancements would yield a more comprehensive approach to balancing safety and performance in shared autonomy environments.

VIII. Conclusion

We proposed an HCSF that significantly enhances system safety while preserving human agency in human–AI shared autonomy settings. Our HCSF builds upon a neural safety value function which is learned scalably through black-box interactions via model-free RL-based HJ reachability analysis. At deployment, we used this value function to enforce a novel Q-CBF safety constraint, which does not require any knowledge of the system dynamics for safety monitoring and intervention. These properties enabled both the synthesis and deployment of our HCSF in Assetto Corsa—a high-fidelity black-box car racing simulator—and make our method the first safety filter applied to high-dimensional, black-box shared autonomy systems involving human operators. Through an extensive in-person user study, we validated two hypotheses using both trajectory data and user responses:

- **H1**: Relative to unassisted driving, our proposed HCSF improves **safety** and user **satisfaction** without compromising human **agency** and **comfort**.
- **H2**: Compared to a conventional safety filter, our proposed HCSF improves human **agency**, **comfort**, and **satisfaction** without compromising **robustness**.

We envision our HCSF being a vital component for a broad class of shared autonomy systems, including advanced driver assistance systems (ADAS) in high-performance driving, thanks to our HCSF's model-free and scalable design. Unlike conventional safety filters—where abrupt interventions that do not take into account the human operator's actions can cause discomfort or automation surprise—our HCSF fosters trustworthy, robust, and agency-preserving human–robot collaboration. Future work will extend our HCSF to time-critical and performance-oriented tasks focusing on the interplay between safety and strategic decision-making, and address the potential over-reliance on safety filters.

References

[1] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs *et al.*, "Outracing champion gran turismo drivers with deep reinforcement learning," *Nature*, vol. 602, no. 7896, pp. 223–228, 2022.

[2] J. DeCastro, A. Silva, D. Gopinath, E. Sumner, T. M. Balch, L. Dees, and G. Rosman, "Dreaming to assist: Learning to align with human objectives for shared control in high-speed racing," in *Proceedings of the 8th Annual Conference on Robot Learning (CoRL)*, 2024.

[3] D. Gopinath, X. Cui, J. DeCastro, E. Sumner, J. Costa, H. Yasuda, A. Morgan, L. Dees, S. Chau, J. Leonard *et al.*, "Computational teaching for driving via multi-task imitation learning," *arXiv preprint arXiv:2410.01608*, 2024.

[4] N. B. Sarter and D. D. Woods, "Team play with a powerful and independent agent: Operational experiences and automation surprises on the Airbus A-320," vol. 39, no. 4, pp. 553–569.

[5] G. A. Jamieson, G. Skraaning, and J. Joe, "The B737 MAX 8 accidents as operational experiences with automation transparency," vol. 52, no. 4, pp. 794–797.

[6] O. Bastani, "Safe reinforcement learning with nonlinear dynamics via model predictive shielding," in *2021 American control conference (ACC)*. IEEE, 2021, pp. 3488–3494.

[7] K.-C. Hsu, H. Hu, and J. F. Fisac, "The safety filter: A unified view of safety-critical control in autonomous systems," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 7, 2023.

[8] SAE, "Taxonomy & definitions for operational design domain (odd) for driving automation systems," *SAE J3259*, 2021.

[9] L. Fraade-Blanar, M. S. Blumenthal, J. M. Anderson, and N. Kalra, *Measuring Automated Vehicle Safety: Forging a Framework*. Santa Monica, CA: RAND Corporation, 2018.

[10] 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.

[11] B. Tearle, K. P. Wabersich, A. Carron, and M. N. Zeilinger, "A predictive safety filter for learning-based racing control," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7635–7642, 2021.

[12] S.-C. Hsu, X. Xu, and A. D. Ames, "Control barrier function based quadratic programs with application to bipedal robotic walking," in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 4542–4548.

[13] A. Agrawal and K. Sreenath, "Discrete control barrier functions for safety-critical control of discrete systems with application to bipedal robot navigation." in *Robotics: Science and Systems*, vol. 13. Cambridge, MA, USA, 2017, pp. 1–10.

[14] D. P. Nguyen, K.-C. Hsu, W. Yu, J. Tan, and J. Fernández Fisac, "Gameplay filters: Robust zero-shot safety through adversarial imagination," in *Proceedings of the 8th Annual Conference on Robot Learning (CoRL)*, 2024.

[15] A. Singletary, A. Swann, Y. Chen, and A. D. Ames, "Onboard safety guarantees for racing drones: High-speed geofencing with control barrier functions," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2897–2904, 2022.

[16] M. Chen, S. L. Herbert, H. Hu, Y. Pu, J. F. Fisac, S. Bansal, S. Han, and C. J. Tomlin, "Fastrack: a modular framework for real-time motion planning and guaranteed safe tracking," *IEEE Transactions on Automatic Control*, vol. 66, no. 12, pp. 5861–5876, 2021.

[17] I. M. Mitchell, A. M. Bayen, and C. J. Tomlin, "A time-dependent hamilton-jacobi formulation of reachable sets for

continuous dynamic games," *IEEE Transactions on automatic control*, vol. 50, no. 7, pp. 947–957, 2005.

[18] S. Bansal, M. Chen, S. Herbert, and C. J. Tomlin, "Hamilton-jacobi reachability: A brief overview and recent advances," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 2242–2253.

[19] J. F. Fisac, N. F. Lugovoy, V. Rubies-Royo, S. Ghosh, and C. J. Tomlin, "Bridging hamilton-jacobi safety analysis and reinforcement learning," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8550–8556.

[20] A. Robey, H. Hu, L. Lindemann, H. Zhang, D. V. Dimarogonas, S. Tu, and N. Matni, "Learning control barrier functions from expert demonstrations," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 3717–3724.

[21] S. Bansal and C. J. Tomlin, "Deepreach: A deep learning approach to high-dimensional reachability," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1817–1824.

[22] K.-C. Hsu, D. P. Nguyen, and J. F. Fisac, "ISAACS: Iterative soft adversarial actor-critic for safety," in *Proceedings of the 5th Annual Learning for Dynamics and Control Conference (L4DC)*. PMLR, 2023, pp. 90–103.

[23] J. Wang, H. Hu, D. P. Nguyen, and J. F. Fisac, "MAGICS: Adversarial rl with minimax actors guided by implicit critic stackelberg for convergent neural synthesis of robot safety," in *Proceedings of the 16th Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2024.

[24] T. He, C. Zhang, W. Xiao, G. He, C. Liu, and G. Shi, "Agile but safe: Learning collision-free high-speed legged locomotion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[25] H. Hu, Z. Zhang, K. Nakamura, A. Bajcsy, and J. Fernández Fisac, "Deception game: Closing the safety-learning loop in interactive robot autonomy," in *Proceedings of the 7th Annual Conference on Robot Learning (CoRL)*, 2023.

[26] K. Leung, E. Schmerling, M. Zhang, M. Chen, J. Talbot, J. C. Gerdes, and M. Pavone, "On infusing reachability-based safety assurance within planning frameworks for human–robot vehicle interactions," *The International Journal of Robotics Research*, vol. 39, no. 10-11, pp. 1326–1345, 2020.

[27] H. Hu, K. Nakamura, and J. F. Fisac, "Sharp: Shielding-aware robust planning for safe and efficient human-robot interaction," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5591–5598, 2022.

[28] G. P. Brat, H. Yu, E. Atkins, P. Sharma, D. Cofer, M. Durling, B. Meng, C. Alexander, S. Borgyos, C. Fan *et al.*, "Autonomy verification & validation roadmap and vision 2045," Tech. Rep., 2023.

[29] M. Chen and C. J. Tomlin, "Hamilton–jacobi reachability: Some recent theoretical advances and applications in unmanned airspace management," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. 1, pp. 333–358, 2018.

[30] E. Squires, P. Pierpaoli, and M. Egerstedt, "Constructive barrier certificates with applications to fixed-wing aircraft collision avoidance," in *2018 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2018, pp. 1656–1661.

[31] D. D. Oh, D. Lee, and H. J. Kim, "Safety-critical control under multiple state and input constraints and application to fixed-wing uav," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 1748–1755.

[32] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3861–3876, 2016.

[33] L. Lindemann, H. Hu, A. Robey, H. Zhang, D. Dimarogonas, S. Tu, and N. Matni, "Learning hybrid control barrier functions from data," in *Conference on robot learning*. PMLR, 2020, pp. 1351–1370.

[34] S. Li and O. Bastani, "Robust model predictive shielding for safe reinforcement learning with stochastic dynamics," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 7166–7172.

[35] K. P. Wabersich and M. N. Zeilinger, "A predictive safety filter for learning-based control of constrained nonlinear dynamical systems," *Automatica*, vol. 129, p. 109597, 2021.

[36] Y. Chow, O. Nachum, E. Duéñez-Guzmán, and M. Ghavamzadeh, "A lyapunov-based approach to safe reinforcement learning," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 8092–8101.

[37] K.-C. Hsu, V. Rubies-Royo, C. J. Tomlin, and J. F. Fisac, "Safety and liveness guarantees through reach-avoid reinforcement learning," in *Proceedings of Robotics: Science and Systems*, Held Virtually, July 2021.

[38] H. Hu, M. Fazlyab, M. Morari, and G. J. Pappas, "Reach-sdp: Reachability analysis of closed-loop systems with neural network controllers via semidefinite programming," in *2020 59th IEEE conference on decision and control (CDC)*. IEEE, 2020, pp. 5929–5934.

[39] M. Everett, G. Habibi, C. Sun, and J. P. How, "Reachability analysis of neural feedback loops," *IEEE Access*, vol. 9, pp. 163 938–163 953, 2021.

[40] O. Gates, M. Newton, and K. Gatsis, "Scalable forward reachability analysis of multi-agent systems with neural network controllers," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 67–72.

[41] A. Lin and S. Bansal, "Verification of neural reachable tubes via scenario optimization and conformal prediction," in *6th Annual Learning for Dynamics & Control Conference*. PMLR, 2024, pp. 719–731.

[42] J. Fisac, A. Bajcsy, S. Herbert, D. Fridovich-Keil, S. Wang, C. Tomlin, and A. Dragan, "Probabilistically safe robot planning with confidence-based human predictions," *Robotics: Science and Systems XIV*, 2018.

[43] R. Tian, L. Sun, A. Bajcsy, M. Tomizuka, and A. D. Dragan, "Safety assurances for human-robot interaction via confidence-aware game-theoretic human models," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 229–11 235.

[44] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2010, pp. 797–803.

[45] H. Hu, D. Isele, S. Bae, and J. F. Fisac, "Active uncertainty reduction for safe and efficient interaction planning: A shielding-aware dual control approach," *The International Journal of Robotics Research*, vol. 43, no. 9, pp. 1382–1408, 2024.

[46] A. Bajcsy, A. Siththaranjan, C. J. Tomlin, and A. D. Dragan, "Analyzing human models that adapt online," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2754–2760.

[47] R. Pandya, C. Liu, and A. Bajcsy, "Robots that learn to safely influence via prediction-informed reach-avoid dynamic games," *arXiv preprint arXiv:2409.12153*, 2024.

[48] A. Broad, T. Murphey, and B. Argall, "Highly parallelized data-driven mpc for minimal intervention shared control," in *Proceedings of Robotics: Science and Systems*, 2019.

[49] C. Schaff and M. R. Walter, "Residual policy learning for shared autonomy," in *Proceedings of Robotics: Science and Systems*, 2020.

[50] S. Reddy, A. D. Dragan, and S. Levine, "Shared autonomy via deep reinforcement learning," in *Proceedings of Robotics: Science and Systems*, 2018.

[51] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep rein-

forcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

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

[53] L. Chen, S. Manuel, J. Delgado, J. Subotsis, and P. Tylkin, "Learn thy enemy: Online, task-aware opponent modeling in autonomous racing," in *Symposium on Machine Learning for Autonomous Driving*, 2023.

[54] J. Lidard, H. Hu, A. Hancock, Z. Zhang, A. G. Contreras, V. Modi, J. DeCastro, D. Gopinath, G. Rosman, N. Leonard, M. Santos, and J. F. Fisac, "Blending data-driven priors in dynamic games," in *Proceedings of Robotics: Science and Systems*, 2024.

[55] H. Hu, J. DeCastro, D. Gopinath, G. Rosman, N. E. Leonard, and J. F. Fisac, "Think deep and fast: Learning neural nonlinear opinion dynamics from inverse dynamic games for split-second interactions," *2025 International Conference on Robotics and Automation (ICRA) (to appear)*, 2024.

[56] B. Chen, J. Francis, J. Oh, E. Nyberg, and S. L. Herbert, "Safe autonomous racing via approximate reachability on ego-vision," *arXiv preprint arXiv:2110.07699*, 2021.

[57] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[58] W. S. Cortez, X. Tan, and D. V. Dimarogonas, "A robust, multiple control barrier function framework for input constrained systems," *IEEE Control Systems Letters*, vol. 6, pp. 1742–1747, 2021.

[59] C. Dawson, Z. Qin, S. Gao, and C. Fan, "Safe nonlinear control using robust neural lyapunov-barrier functions," in *Conference on Robot Learning*. PMLR, 2022, pp. 1724–1735.

[60] O. So, Z. Serlin, M. Mann, J. Gonzales, K. Rutledge, N. Roy, and C. Fan, "How to train your neural control barrier function: Learning safety filters for complex input-constrained systems," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 532–11 539.

[61] S. Zhang, O. So, K. Garg, and C. Fan, "Gcbf+: A neural graph control barrier function framework for distributed safe multi-agent control," *IEEE Transactions on Robotics*, vol. PP, no. 99, pp. 1–20, 2025.

[62] W. Xiao, T.-H. Wang, R. Hasani, M. Chahine, A. Amini, X. Li, and D. Rus, "Barriernet: Differentiable control barrier functions for learning of safe robot control," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 2289–2307, 2023.

[63] Z. Li, J. Zeng, A. Thirugnanam, and K. Sreenath, "Bridging model-based safety and model-free reinforcement learning through system identification of low dimensional linear models," in *Proceedings of Robotics: Science and Systems*, 2022.

[64] T. G. Molnar, R. K. Cosner, A. W. Singletary, W. Ubellacker, and A. D. Ames, "Model-free safety-critical control for robotic systems," *IEEE robotics and automation letters*, vol. 7, no. 2, pp. 944–951, 2021.

[65] M. H. Cohen, T. G. Molnar, and A. D. Ames, "Safety-critical control for autonomous systems: Control barrier functions via reduced-order models," *Annual Reviews in Control*, vol. 57, p. 100947, 2024.

[66] E. Squires, R. Konda, S. Coogan, and M. Egerstedt, "Model free barrier functions via implicit evading maneuvers," *arXiv preprint arXiv:2107.12871*, 2021.

[67] S. S. Kumar, Q. Lin, and J. Dolan, "Latentcbf: A control barrier function in latent space for safe control," 2024.

[68] Y. Yang, Y. Jiang, Y. Liu, J. Chen, and S. E. Li, "Model-free safe reinforcement learning through neural barrier certificate," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1295–1302, 2023.

[69] K. He, S. Shi, T. v. d. Boom, and B. De Schutter, "State-action control barrier functions: Imposing safety on learning-based control with low online computational costs," *arXiv preprint arXiv:2312.11255*, 2023.

[70] J. Borquez, K. Chakraborty, H. Wang, and S. Bansal, "On safety and liveness filtering using hamilton-jacobi reachability analysis," *IEEE Transactions on Robotics*, 2024.

[71] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[72] A. Remonda, N. Hansen, A. Raji, N. Musiu, M. Bertogna, E. E. Veas, and X. Wang, "A simulation benchmark for autonomous racing with large-scale human data," in *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[73] J. Roozendaal, E. Johansson, J. d. Winter, D. Abbink, and S. Petermeijer, "Haptic lane-keeping assistance for truck driving: a test track study," *Human factors*, vol. 63, no. 8, pp. 1380–1395, 2021.

[74] J. W. Moore, "What is the sense of agency and why does it matter?" *Frontiers in psychology*, vol. 7, p. 1272, 2016.

[75] N. Braun, S. Debener, N. Spychala, E. Bongartz, P. Sörös, H. H. Müller, and A. Philipsen, "The senses of agency and ownership: a review," *Frontiers in psychology*, vol. 9, p. 535, 2018.

[76] J. D. Loehr, "The sense of agency in joint action: An integrative review," *Psychonomic Bulletin & Review*, vol. 29, no. 4, pp. 1089–1117, 2022.

[77] F. Mueller, N. Semertzidis, J. Andres, J. Marshall, S. Benford, X. Li, L. Matjeka, and Y. Mehta, "Toward understanding the design of intertwined human–computer integrations," *ACM Transactions on Computer-Human Interaction*, vol. 30, no. 5, pp. 1–45, 2023.

[78] T. J. Prescott, K. Vogeley, and A. Wykowska, "Understanding the sense of self through robotics," *Science robotics*, vol. 9, no. 95, p. eadn2733, 2024.

[79] M. A. Collier, R. Narayan, and H. Admoni, "The sense of agency in assistive robotics using shared autonomy," *arXiv preprint arXiv:2501.07462*, 2025.

[80] C. Sohn, J. Andert, and R. N. N. Manfouo, "A driveability study on automated longitudinal vehicle control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3273–3280, 2019.

[81] M. Werling, S. Kammel, J. Ziegler, and L. Gröll, "Optimal trajectories for time-critical street scenarios using discretized terminal manifolds," *The International Journal of Robotics Research*, vol. 31, no. 3, pp. 346–359, 2012.

[82] H. Wang, B. Liu, X. Ping, and Q. An, "Path tracking control for autonomous vehicles based on an improved mpc," *IEEE access*, vol. 7, pp. 161 064–161 073, 2019.

[83] 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.

APPENDIX

A. Proofs of Theoretical Results

In this section, we present the proofs for Proposition 1 and Proposition 2. We first state two key lemmas.

Lemma 1. *For all $x \in \mathcal{X}$, there exists an action $u \in \mathcal{U}$ such that $\mathcal{Q}(x,u) \geq \gamma V(x), \quad \forall \gamma \in [0,1)$*

Proof: From (6), $\exists u \in \mathcal{U}$ such that $\mathcal{Q}(x,u) = V(x)$. One such action is from the best-effort fallback policy, $u^{\blacklozenge}(x) := \text{argmax}_{u \in \mathcal{U}} \mathcal{Q}(x,u)$. For any $\gamma \in [0,1)$ we have $\gamma V(x) \leq V(x)$. Therefore, there always exists an action $u \in \mathcal{U}$ such that $\mathcal{Q}(x,u) \geq \gamma V(x), \quad \forall \gamma \in [0,1)$. ■

Lemma 2. *For all $x \in \mathcal{X}$, $\forall \gamma \in [0,1)$, $\mathcal{Q}(x,u) \geq \gamma V(x)$ is equivalent to $V(f(x,u)) - V(x) \geq -\alpha V(x), \quad$ where $\alpha = 1-\gamma$.*

Proof: From Lemma 1 we have $\mathcal{Q}(x,u) \geq \gamma V(x)\ \forall \gamma \in [0,1)$. Subtracting $V(x)$ from both sides results in $\mathcal{Q}(x,u) - V(x) \geq (\gamma - 1)V(x)$. Using the definition of the $\mathcal{Q}$-function from (6) we substitute for $\mathcal{Q}(x,u))$:

$$\min\{g(x), \max_{u' \in \mathcal{U}} \mathcal{Q}(f(x,u),u')\} - V(x) \geq (\gamma - 1)V(x)$$

This inequality holds if and only if both of the following conditions are satisfied:

$$g(x) - V(x) \geq (\gamma - 1)V(x), \quad \text{(Condition 1)}$$
$$\max_{u' \in \mathcal{U}} \mathcal{Q}(f(x,u)),u') - V(x) \geq (\gamma - 1)V(x) \quad \text{(Condition 2)}$$

Analyzing Condition 1: from the safe set definition, $V(x) \leq g(x)$ for all $x \in \mathcal{X}$. Thus $g(x) - V(x) \geq 0$. Since $\gamma \in [0,1)$ implies $\gamma - 1 < 0$, it follows that $g(x) - V(x)) \geq (\gamma - 1)V(x)$. Analyzing Condition 2: from (4), $V(f(x,u)) = \max_{u' \in \mathcal{U}} \mathcal{Q}(f(x,u),u')$. Substituting this in Condition 2 we have, $V(f(x,u)) - V(x) \geq (\gamma - 1)V(x)$. Hence, both conditions hold for all $x \in \mathcal{X}$, $u \in \mathcal{U}$, and $\gamma \in [0,1)$. Therefore for $\alpha = 1 - \gamma$,

$$\mathcal{Q}(x,u) \geq \gamma V(x) \iff V(f(x,u)) - V(x) \geq -\alpha V(x)$$

■

We are now ready to prove Proposition 1 and Proposition 2. We restate them here for convenience.

Proposition (Restatement of Proposition 1)**.** *The safety value function $V(x) : \mathcal{X} \to \mathbb{R}$, which is a fixed-point solution of the safety Bellman equation (4), is a valid DCBF as defined in Definition 1 and Remark 1. The corresponding DCBF constraint is:*

$$\mathcal{Q}(x,u) \geq \gamma V(x), \tag{15}$$

where $\gamma \in [0,1)$.

Proof: To prove that $V(x)$ is a valid DCBF, it is sufficient to show that for all $x \in \mathcal{S}$, there exists an action $u \in \mathcal{U}$ such that:

$$V(f(x,u)) - V(x) \geq -\alpha V(x), \tag{16}$$

where $\alpha \in (0,1)$. By Lemma 1 and Lemma 2, the inequality $\mathcal{Q}(x,u) \geq \gamma V(x)$ is equivalent to the DCBF condition above, with $\alpha = 1 - \gamma$. This equivalence establishes that the $\mathcal{Q}$-function directly defines the DCBF constraint in a model-free manner, replacing the conventional dynamics-based constraint. ■

Proposition (Restatement of Proposition 2)**.** *The optimization problem in (11) is recursively feasible for $\forall \gamma \in [0,1)$, given the initial state $x \in \mathcal{S}$.*

Proof: By Lemma 1, for all $x \in \mathcal{S}$ and $\gamma \in [0,1)$, $\exists u \in \mathcal{U}$ such that $\mathcal{Q}(x,u) \geq \gamma V(x) \implies \mathcal{Q}(x,u) \geq 0$. From (4), this choice of u ensures $V(f(x,u)) \geq 0 \implies f(x,u) \in \mathcal{S}$, meaning the system remains within the safe set $\mathcal{S}$ at the next timestep. Since this reasoning applies recursively for all timesteps, the optimization problem (11) is feasible for $\forall \gamma \in [0,1)$. ■

B. Additional Environment Details

1) Observation: We use a 133-dimensional observation vector for both neural synthesis of safety filters and for training the warmup policies. Table IV lists the elements of this observation vector, their corresponding dimensionalities, and the number of past timesteps (i.e., "stacked" frames) used:

TABLE IV: Hyperparameters for the Warmup & Initialization Phases

Observation	Dimensionality	Past Timesteps
Ego vehicle speed	1	4
Gap to the reference path	1	4
Force feedback	1	4
RPM	1	4
Acceleration	2	4
Gear	1	4
Angular velocity	1	4
Local velocity	2	4
Slip Angle	4	4
Distance to track boundary	11	4
Out-of-track	1	1
Look ahead curvature	12	1
Control input	3	4
Ego vehicle heading	1	1
Ego vehicle yaw	1	1
Opponent distance	1	1
Opponent direction	1	1
Opponent speed	1	1
Opponent heading	1	1
Opponent yaw	1	1
Opponent brake	1	1

2) Action: The actions represent increments to the control inputs over a single timestep rather than absolute control values. This incremental representation mitigates oscillations or chattering in the control signals [72].

C. Extended Implementation Details

In this section, we provide additional details about our training methodology. Refer to Table V for the hyperparameter values mentioned in this subsection.

1) Warmup and Initialization: The training pipeline for HCSF is designed to promote robustness and improve learning efficiency in the high-fidelity simulation environment of AC. It operates in three distinct phases per episode—warmup, initialization, and training—each ensuring the ego agent encounters a broad spectrum of scenarios, including failure-prone states that yield the most valuable data for training the safety filter.

Warmup Phase. In AC, each reset places the ego agent in a trivial, stationary configuration on the reference path, with its heading tangent to the path. Because it remains in the safe set if it does not move, this state provides little value for learning the $\mathcal{Q}$-function or the best-effort fallback policy—particularly regarding near-failure conditions. To address this, we introduce a *warmup phase* that accelerates the agent to higher speeds and initiates interactions with the track or opponents before formal training begins.

During warmup, the agent follows either an overtaking policy $\pi_{\text{over}}^{\text{warmup}}$ or a nominal policy $\pi_{\text{nom}}^{\text{warmup}}$ for a duration sampled from $[0, T^{\text{warmup,max}}]$. These two policies are chosen with probability $P_{\text{over}}^{\text{warmup}}$ and its complement, respectively, exposing the agent to both opponent-aware and opponent-agnostic driving scenarios. This warmup procedure transitions the agent away from trivial reset conditions and into more realistic states where the safety filter's intervention is most meaningful. For additional details regarding the training of the warmup policies, see the next subsection.

To diversify starting states, the warmup phase may terminate early when the ego agent enters failure-prone regions, with probability $P_{\text{oppo}}^{\text{warmup}}$ if it is within a distance sampled from $[d_{\text{oppo}}^{\text{warmup,min}}, d_{\text{oppo}}^{\text{warmup,max}}]$ of the nearest opponent. This increases the frequency of interactions with opponents and helps in learning collision avoidance strategies. Additionally, heavy braking can trigger early termination: at the start of each episode, a gating probability $P_{\text{brake, epi}}^{\text{warmup}}$ determines whether braking is considered for termination. If allowed, termination occurs with probability $P_{\text{brake, step}}^{\text{warmup}}$ at each timestep when the braking input exceeds $u_{\text{brake}}^{\text{warmup}}$ and speed is above v^{warmup}. Employing these probabilities sequentially ensures that heavy braking only leads to termination under conditions resembling realistic failure scenarios. This phased termination approach helps the agent reach states critical for learning an effective safety filter.

Initialization Phase. While the warmup phase introduces motion and some variability in initial conditions, it relies primarily on well-trained policies like $\pi_{\text{over}}^{\text{warmup}}$, which tend to stay on the reference path and avoid collisions. As a result, it rarely induces suboptimal driving behaviors or pushes the ego agent into failure-prone scenarios. This limitation is problematic because effective learning of the $\mathcal{Q}$-function and the best-effort fallback policy demands spanning the full state space $\mathcal{X}$ and control space $\mathcal{U}$ [19].

To address this, we introduce the *initialization phase*, which systematically exposes the ego agent to failure-prone regions of $\mathcal{X}$ and $\mathcal{U}$. At each timestep, the agent selects an action and queries the $\mathcal{Q}$-function to check if the resulting state–action pair falls below $\mathcal{Q}_{\text{term}}^{\text{init}}$, indicating a potentially hazardous scenario. With probability $P_{\text{term}}^{\text{init}}$, the initialization phase terminates, and the ego vehicle can start the training phase in this "dangerous" state where safety intervention is most critical. Because human drivers often miss braking points before corners and maintain throttle when braking is required, the initialization phase also simulates a *full-throttle mode* with probability $P_{\text{FT}}^{\text{init}}$. We define $\mathcal{U}_{\text{FT}} = \{(u_{\text{steer}}, u_{\text{throttle}}, u_{\text{brake}}) \mid u_{\text{steer}} \in [-1, 1], u_{\text{throttle}} = 1, u_{\text{brake}} = -1\}$, representing high-speed or aggressive control signals aimed at inducing failure-prone conditions stemming from late braking.

Three initialization schemes—*adversarial*, *random*, and *mixed*—provide distinct ways to sample actions:

- **Adversarial Initialization:** Chosen with probability $P_{\text{adv}}^{\text{init}}$. Actions come from either $\mathcal{U}_{\text{FT}}$ (if full-throttle mode is engaged) or $\mathcal{U}$. The $\mathcal{Q}$-function identifies the action with the smallest $\mathcal{Q}$-value (the "adversarial" action). If its value lies below $\mathcal{Q}_{\text{term}}^{\text{init}}$, the phase terminates with probability $P_{\text{term}}^{\text{init}}$. This repeats until termination.
- **Random Initialization:** Chosen with probability $P_{\text{rand}}^{\text{init}}$. Actions are sampled randomly from $\mathcal{U}_{\text{FT}}$ or $\mathcal{U}$, without querying $\mathcal{Q}$. If the $\mathcal{Q}$-value for a sampled action is below $\mathcal{Q}_{\text{term}}^{\text{init}}$, termination occurs with probability $P_{\text{term}}^{\text{init}}$.
- **Mixed Initialization:** Chosen with probability $P_{\text{mix}}^{\text{init}}$. At each timestep, the agent switches between adversarial or random strategies with probability 0.5. Termination follows the same rule, triggered by $P_{\text{term}}^{\text{init}}$ whenever the $\mathcal{Q}$-value is below $\mathcal{Q}_{\text{term}}^{\text{init}}$.

This approach helps the system generalize to real deployment scenarios, where human operators may have varying skill levels and intentions, leading to unpredictable behaviors and frequent exposure to failure-prone conditions.

2) Training the Warmup Policies: We employ two different warmup policies during the warmup phase: a *nominal* policy and an *overtaking* policy. The nominal policy, trained using the reward function from [72], seeks the fastest lap times without considering opponents. In contrast, the overtaking policy accounts for the nearest opponent, rewarding successful overtakes and penalizing collisions, although it does not guarantee collision avoidance. The nominal policy is trained using the following reward function:

$$r_{\text{nom}} = \frac{v}{c_2}\left(1 - c_1 \cdot d_{\text{gap}}\right), \tag{17}$$

where v denotes the speed of the ego vehicle, d_{gap} is the ℓ_2-distance to the reference path, and c_1 and c_2 are design parameters.

For the overtaking policy, we encourage overtaking maneuvers using a term inspired by [1]:

$$r_{\text{over},1} = \mathrm{I}\{d_{\text{oppo}} < d_{\text{oppo}}^{\text{over}}\} \cdot c_3 \cdot (\Delta\text{NSP}_{t-1} - \Delta\text{NSP}_t), \tag{18}$$

where "oppo" indicates the nearest opponent, $\mathrm{I}\{\cdot\}$ is an indicator function that equals 1 if $d_{\text{oppo}} < d_{\text{oppo}}^{\text{over}}$ and 0 otherwise, c_3 is a design parameter, and $\Delta\text{NSP}_t := \text{NSP}_{\text{oppo},t} - \text{NSP}_{\text{ego},t}$. Here, "NSP" (normalized spline position) measures progression along the track, taking values in $[0, 1)$. By construction, $r_{\text{over},1} > 0$ if and only if either the ego vehicle is behind

TABLE V: Hyperparameters for the Warmup & Initialization Phases

Hyperparameter	Value	Description
$T^{\text{warmup,max}}$	25 s	Max duration of warmup phase.
$P^{\text{warmup}}_{\text{over}}$	0.6	Probability of using overtaking policy during warmup.
$P^{\text{warmup}}_{\text{oppo}}$	0.25	Probability of early termination if near an opponent.
$d^{\text{warmup,min}}_{\text{oppo}}$	6 m	Min distance threshold for opponent proximity.
$d^{\text{warmup,max}}_{\text{oppo}}$	36 m	Max distance threshold for opponent proximity.
$P^{\text{warmup}}_{\text{brake, epi}}$	0.4	Probability that heavy braking triggers warmup termination.
$P^{\text{warmup}}_{\text{brake, step}}$	0.25	Probability of termination on each timestep of heavy braking.
$u^{\text{warmup}}_{\text{brake}}$	0.6	Threshold for heavy braking input.
v^{warmup}	40 m/s	Speed threshold under which braking triggers termination.
$P^{\text{init}}_{\text{term}}$	0.2	Probability of ending initialization once $\mathcal{Q}$-value falls below threshold.
$\mathcal{Q}^{\text{init}}_{\text{term}}$	2	$\mathcal{Q}$-value threshold for indicating dangerous scenarios.
$P^{\text{init}}_{\text{FT}}$	0.4	Probability of enabling full-throttle mode.
$P^{\text{init}}_{\text{adv}}$	0.3	Probability of adversarial initialization.
$P^{\text{init}}_{\text{rand}}$	0.3	Probability of random initialization.
$P^{\text{init}}_{\text{mix}}$	0.4	Probability of mixed initialization.
c_1	$1/12$	Design parameter for training the nominal warmup policy.
c_2	300	Design parameter for training the nominal warmup policy.
$d^{\text{over}}_{\text{oppo}}$	100 m	Distance threshold for overtaking rewards.
c_3	600	Design parameter for training the overtaking policy.

the opponent and is closing in, or the ego vehicle is ahead of the opponent and is pulling away. Overtaking is further encouraged with an additional term:

$$r_{\text{over},2} = \mathrm{I}\{r_{\text{over},1} > 0\} \cdot \frac{c_1}{c_2} \cdot v \cdot d_{\text{gap}}, \tag{19}$$

which activates only if the ego vehicle is closing in on the opponent from behind or pulling away in front. In effect, this negates the penalty term for deviating from the reference path in (17), allowing the ego agent to set up or complete an overtake without strictly adhering to the reference path. The entire reward function for training the overtaking policy is as follows:

$$r_{\text{over}} = r_{\text{nom}} + \max(r_{\text{over},1}, 0) + r_{\text{over},2}. \tag{20}$$

We also note that, when training the overtaking policy, the ego vehicle is heavily penalized for colliding with the opponent through immediate episode termination.

3) Training Details: Following [19, 22], we approximate the state–action safety value function $\mathcal{Q}(\cdot,\cdot)$ using a neural network with parameters ϕ. We also parameterize the best-effort fallback policy as a separate neural network with parameters θ. Our goal is for $\mathcal{Q}_\phi$ to approximate a fixed-point solution of the time-discounted safety Bellman equation, defined as:

$$V_\phi(x) = (1-\gamma_{\text{ENV}})g(x) + \gamma_{\text{ENV}} \min\{g(x), \max_u Q_\phi(x,u))\}. \tag{21}$$

To remain compatible with standard reinforcement learning methods such as SAC [71], we train the $\mathcal{Q}$-network to minimize the Bellman residual:

$$L(\phi) = \mathbb{E}^{\mathcal{B},\pi}[(Q_\phi(x,u) - y)^2], \tag{22}$$

where $y := (1-\gamma_{\text{ENV}})\, g' \;+\; \gamma_{\text{ENV}} \min\{g', Q_\phi(x', u')\}$, and $Q_{\phi'}$ is the target $\mathcal{Q}$-network. We then update π_θ using the following policy gradient:

$$L(\theta) := \mathbb{E}^{\mathcal{B},\pi}[-Q_\phi(x,u) + \alpha \log \pi_\theta(u|x)]. \tag{23}$$

(22) and (23) are updated in an alternating manner by sampling from $\mathcal{B}$. We employ double $\mathcal{Q}$-learning and update the $\mathcal{Q}$-networks with a delay. Table VI summarizes our hyperparameters. The complete HCSF training pseudocode is given in Algorithm 1.

TABLE VI: Hyperparameters for SAC Training

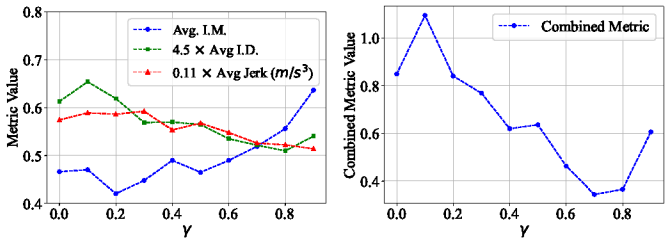
Hyperparameter	Value	Description
η	3×10^{-4}	Actor and critic learning rate.
γ_{ENV}	0.992	Discount factor.
$\lvert\mathcal{B}\rvert$	2×10^7	Replay buffer size.
Batch Size	256	Training batch size.
τ	0.005	Target network update rate.
α	Learned	Entropy temperature.
N_{UTD}	1	Gradient steps per environment step.

4) Solving the OCP: We solve the OCP (11) by sampling 2000 candidate actions on the line that connects the human's action $u^{\text{human}}(x)$ and the best-effort fallback policy's action $u^{\blacklozenge}(x)$ in the control space. Among the candidate actions, we select the one that minimizes the deviation from $u^{\text{human}}(x)$ while satisfying (11b) (note that $u^{\blacklozenge}(x)$ always satisfies (11b)). This is because solely relying on the learned $\mathcal{Q}$-function to solve the OCP (11), such as sampling in the entire control space and searching for the candidate that minimizes the deviation from $u^{\text{human}}(x)$ while satisfying (11b), led to unsatisfactory safety metrics in practice. We compensate for the neural approximation error in the $\mathcal{Q}$-function by additionally leveraging the information encoded in the best-effort fallback policy $u^{\blacklozenge}(x)$ [22], thereby achieving near-zero failures throughout the experiment. Optimization remains computationally efficient, as candidate actions are rapidly propagated through the Q-network using GPU-accelerated tensor computations. The pseudocode for HCSF optimization is given in Algorithm 2.

5) Choice of Design Parameters: γ is a design parameter in our HCSF formulation (11) that does *not* affect the *safety guarantees* but does influence the shared autonomy system's *behavior*. In related works [70], γ is tuned for objectives that are not directly related to safety. In our HCSF framework, we tune γ to modulate human agency, comfort, and satisfaction. Based on a preliminary trial among the authors, we determined that $\gamma = 0.7$ gives a strong sense of agency and satisfaction while feeling smooth, and we used this value throughout the user study.

We further analyze γ's effect on agency and comfort via an ablation study in which the overtaking policy $\pi_{\text{over}}^{\text{warmup}}$ that was trained using the reward function (20) raced the ego vehicle against an opponent in lieu of human drivers. We tested 10 instantiations of our HCSF with different γ values ranging from 0 to 0.9, and each instantiation was tested for 10 minutes. We note that this setting is identical to driving session 2 of our user study. The average I.M., which is a proxy for the agency metric, and average I.D. and jerk, which are proxies for comfort, for each HCSF instantiation are shown in Fig. 15a. Smaller I.M. indicates better agency, while smaller I.D. and jerk indicate better comfort. We observe that as γ becomes larger, proxies for comfort improve while the proxy for agency degrades. This is coherent with the understanding of $\alpha = 1-\gamma$ in conventional DCBF settings (9b).

We define the *Combined Metric* as the uniformly weighted sum of I.M., I.D., and jerk after linearly normalizing them to be between 0 and 1. As shown in Fig. 15b, we found $\gamma = 0.7$ to be a sweet spot between agency and comfort, a result coinciding with our preliminary trial.



(a) Average I.M., I.D., and jerk. **(b)** Combined Metric on different γ.

Fig. 15: HCSF with $\gamma = 0.7$ strikes a good balance between the proxies for agency and comfort.

Our framework can also integrate approaches like Parametric CBFs [83] that co-optimize γ to further enhance human agency, comfort, and satisfaction without compromising safety. We acknowledge the relevance of our work to the smooth least restrictive safety filter [70], which is an instantiation of our HCSF using $\gamma = 0$ and does not necessarily strike a good balance between agency and comfort, as can be seen in Fig. 15.

6) Wall-Clock Training Time: The three-week training time stems primarily from the AC simulator's real-time constraint, with a majority of the time spent on warmup and initialization phases. During the training phase, which begins after the warmup and initialization phases, we sample transitions from a single environment and append them to $\mathcal{B}$.

D. Extended Results

1) Cronbach's Alpha Test: Table VII summarizes the Cronbach's alpha values obtained for each of the eight metrics. Agency (0.716), comfort (0.775), and satisfaction (0.710) exceed the commonly cited threshold of 0.70, indicating acceptable internal consistency. Robustness (0.673) and trustworthiness (0.683) fall just below 0.70 and can be considered borderline acceptable, which is reasonable given that each metric was assessed using only two items. Interpretability (0.562) and competence (0.540) are borderline, suggesting that the negated items may not perfectly capture the reverse of their affirmative counterparts. Predictability (0.175) stands out as poor, indicating a potential mismatch between its affirmative and negated statements or participants' interpretations.

Despite this variability, the four core metrics—robustness, agency, comfort, and satisfaction—either exceed or closely approach the 0.70 threshold, making them sufficiently reliable for supporting our core hypotheses. Hence, we consider the qualitative data for these metrics to be reliable enough to validate the conclusions drawn from our study.

TABLE VII: Cronbach's Alpha Test Results

Metric	Cronbach's Alpha Value
Robustness	0.673
Agency	0.716
Comfort	0.775
Satisfaction	0.710
Trustworthiness	0.683
Predictability	0.175
Interpretability	0.562
Competence	0.540

2) Racing Performance: The final lap times for each participant in all three sessions—1, 2, and 3—are shown in Fig. 16. We use final lap time as a quantitative performance metric, particularly for session 2 in which participants receive a safety filter (including a placebo). We assume that participants need some time to learn how to cooperate with the assistance. Moreover, because each failure forces the vehicle to reset—causing it to restart from a stationary position—we treat failure incidents as inherently penalized in terms of lap times.

Although no statistical significance was observed, participants who received a safety filter in session 2 showed greater improvements in lap times compared to those in the unassisted group. Notably, the HCSF group in session 2 was the only one to average a sub-three-minute lap time. This advantage likely stems from HCSF providing gradual, smoother interventions rather than the last-minute, large corrections of LRSF. Such preemptive, smooth corrections have been linked to better performance metrics (e.g., lap times) in filter-aware motion planning literature [27, 45].

Although no statistical significance was observed, Fig. 16 suggests a tendency toward over-reliance on safety filters. While the average final lap time of unassisted participants decreased across sessions (likely due to gained driving experience), both LRSF and HCSF groups—despite showing

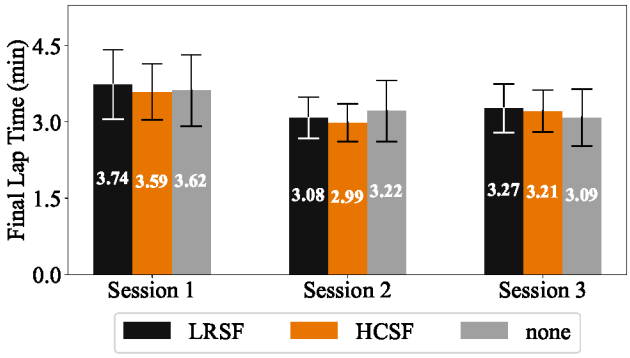


Fig. 16: Average final lap times for each group across sessions 1, 2, and 3.

the shortest times in session 2—had increased lap times in session 3. They became slightly slower (though not significantly) than the unassisted group, potentially because they relied too heavily on the safety filters and thus saw less improvement in driving skill. We anticipate that future HCSF designs, which can foresee a longer time horizon and make *strategic, performance-oriented* decisions beyond mere safety enforcement, may address this issue of over-reliance.

Algorithm 1: HCSF Training

Initialize policy parameters θ, value function parameters ϕ
Initialize replay buffer $\mathcal{D}$
Set target smoothing coefficient τ, discount factor γ_{ENV}, and temperature α
for each training episode **do**
  Observe initial state x_0
  for each environment step **do**
    Select action $u_t \sim \pi_\theta(u_t|x_t)$
    Execute u_t in environment, observe margin g_t, and next state x_{t+1}
    Store transition (x_t, u_t, g_t, x_{t+1}) in replay buffer $\mathcal{B}$
  end for
  for each gradient step $i = 1, \ldots, N_{UTD}$ **do**
    Sample a minibatch of transitions (x_t, u_t, g_t, x_{t+1}) from $\mathcal{B}$
    Compute target $y_t := (1 - \gamma_{\text{ENV}})g_t + \gamma_{\text{ENV}} \min\{g_t, Q_\phi(x', u')\}$
    Update $\mathcal{Q}$-function(s) using (22)
    Update policy using (23)
    Update target $\mathcal{Q}$-function(s): $\phi \leftarrow \tau\phi + (1 - \tau)\phi$
  end for
end for

Algorithm 2: HCSF Execution

for each environment step **do**
  Observe the current state x
  Observe human action $u^{\text{human}}(x)$ and compute the best-effort fallback policy's action $u^{\blacktriangledown}(x)$
  Sample candidate actions on the line that connects $u^{\text{human}}(x)$ and $u^{\blacktriangledown}(x)$
  Select a candidate action u that minimizes (11a) while satisfying (11b).
  Apply the filtered output u to the system
end for