Survey

Manuel Hess, Ben-Micha Piscol*, Christopher Bohn and Sören Hohmann

Safety filters as a means for ensuring functional safety in data-driven control: an overview

Safety Filter als Mittel zur Gewährleistung funktionaler Sicherheit in datengetriebenen Regelungssystemen: Ein Überblick

https://doi.org/10.1515/auto-2025-0104
Received September 30, 2025; accepted April 7, 2026

Abstract: This paper highlights the necessity of provable runtime safety mechanisms for integrating data-driven control methods into safety-critical systems operating in open environments. Since data-driven methods are often limited in providing formal guarantees, we argue for the use of formal methods, such as safety filters, to provide safety assurances. Safety filters denote a class of runtime mechanisms that ensure safety even if the nominal control method does not ensure safety, by constraining the system state to provably safe sets. The focus of this work is on analyzing safety filters from the perspective of functional safety as defined in industrial standards. We demonstrate how safety filters can provably reduce risks associated with hazardous behavior and how they operate as a monitoring and intervention mechanism for data-driven methods.

Keywords: automated systems; data-driven control; functional safety; open environment; safety filter

Zusammenfassung: Diese Arbeit hebt die Notwendigkeit nachweisbarer Sicherheitsmechanismen für die Integration datengetriebener Regelungsverfahren in sicherheitskritische Systeme hervor, die in offenen Umgebungen betrieben werden. Da datengetriebene Methoden häufig nur eingeschränkte formale Garantien liefern können, argumentieren wir für den Einsatz formaler Methoden, wie beispielsweise Safety Filter, zur Gewährleistung der Sicherheit. Safety Filter bezeichnen eine Klasse von Mechanismen, die zur Laufzeit Sicherheit auch dann gewährleisten, wenn die nominale Regelungsmethode dies nicht gewährleistet, indem sie den Systemzustand auf nachweislich sichere Mengen beschränken. Der Fokus dieser Arbeit liegt auf der Analyse von Safety Filtern aus der Perspektive der funktionalen Sicherheit im Kontext industrieller Normen. Es wird gezeigt, wie Safety Filter Risiken im Zusammenhang mit gefährlichem Verhalten nachweisbar reduzieren können und als Mechanismus für Monitoring und Intervention für datengetriebene Verfahren fungieren.

Schlagwörter: Automatisierte Systeme; datengetriebene Regelung; funktionale Sicherheit; offene Umgebung; Safety Filter

1 Introduction

Automation is advancing rapidly in open and dynamic environments across a wide range of application domains [1]–[5]. Open environments are characterized by an unbounded set of possible scenarios, which increases the complexity of control tasks, as many scenarios cannot be fully anticipated during system design. In this context, and with the growing availability of data and computational resources [6], data-driven methods are increasingly employed for planning and control tasks [7], [8]. A key advantage of data-driven methods is their ability to capture a broad spectrum of operating conditions and improve performance beyond classical control [8].

Ensuring safety directly through data-driven methods remains a challenging task [9], [10], leading to safety concerns regarding their use in safety-critical domains [11]–[13]. Data-driven control methods typically rely on models that are at least partially black-box in nature. This is in contrast to the first-principles models usually used to formally analyze closed-loop dynamics in classical

Manuel Hess and Ben-Micha Piscol contributed equally to this work and share first authorship.

***Corresponding author: Ben-Micha Piscol**, Institute of Control Systems (IRS), Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany, E-mail: ben-micha.piscol@kit.edu
Manuel Hess, Christopher Bohn and Sören Hohmann, Institute of Control Systems (IRS), Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany, E-mail: manuel.hess@kit.edu (M. Hess), christopher.bohn@kit.edu (C. Bohn), soeren.hohmann@kit.edu (S. Hohmann)

control theory. In particular, the partially black-box nature of such controllers reduces interpretability and complicates the derivation of rigorous mathematical safety proofs [14], [15].

However, in particular if human lives are at risk, safety is a critical requirement for automated systems, as reflected in numerous industrial standards and regulations on functional safety [16]–[18]. Recent work highlights fundamental challenges in ensuring safety in real-world environments [19]. In the standards on functional safety [16], a central approach to ensure safety in safety-critical systems is the application of a verification and validation (V&V) process. In this context, validation largely relies on systematic testing. However, the complexity and variability of scenarios in open environments, in which data-driven methods are expected to operate, make exhaustive testing increasingly impractical [20]. In particular, the so-called *long tail* of rare but safety-critical scenarios renders exhaustive validation practically infeasible [20].

Recently, safety filters have been proposed in the literature as a promising approach for ensuring safety at runtime instead of relying solely on exhaustive testing and validation [21], [22]. A safety filter continuously monitors the system state and intervenes only if necessary to keep the system state within a provably safe set, while otherwise leaving control to a nominal controller.

However, to the best of our knowledge, it remains unclear how, and to what extent, safety filters contribute to functional safety in accordance with regulatory and industrial standards. Without a fundamental understanding of the role safety filters may play in certification, even the most promising theoretical methods are unlikely to translate into real-world products.

1.1 Related work

Recent surveys and foundational works investigate safety filters and related runtime safety mechanisms [21]–[24]. In particular, unifying perspectives on the formulation and application of safety filters are provided in [21], [22], while foundational contributions establish control-theoretic safety filter concepts such as control barrier functions [24]. Furthermore, connections between learning-based control and formal safety mechanisms are discussed in the context of safe learning and robotics [23].

While these works offer comprehensive methodological overviews and conceptual foundations, their focus lies on technical and algorithmic concepts, architectures, and application domains. The integration of safety filters into established functional safety standards and certification processes is not examined. ISO/IEC TR 5469 [25] highlights the relevance of supervision functions for constraining AI-based systems within defined safety limits, but it does not analyze how specific safety filter approaches relate to functional safety. Consequently, a structured assessment of the role of safety filters in the context of industrial functional safety standards remains open.

1.2 Contribution

In this paper, we adopt a *functional safety perspective* on safety filters, complementing prior works that primarily address theoretical and algorithmic aspects. Our contribution is twofold:

- We provide a survey of functional safety concepts as defined in relevant industrial standards, alongside a survey of safety filters. By presenting both perspectives in a unified notation, we establish a common ground between the two research fields.
- We conceptually relate the requirements and recommendations of relevant industrial standards to the operational principles of safety filters. In doing so, we discuss how safety filters may support the achievement of functional safety objectives. Specifically, we highlight their potential as risk reduction measures, their influence on testing and validation considerations, and their connection to formal methods for ensuring the safety of data-driven control approaches.

1.3 Outline

Section 2 presents an overview of the functional safety landscape, introduces relevant processes and terminology, and examines a representative requirement and a specific recommendation from industrial standards. It further outlines challenges in ensuring functional safety for data-driven methods in open environments. In Section 3, we provide a formal definition of safety for autonomous systems in open environments. Building on this, Section 4 summarizes the concept of safety filters and gives an overview of state-of-the-art methods. The section concludes with a collision-avoidance example that illustrates the behavior of different safety filter methods. Finally, Section 5 relates the requirement and the recommendation introduced in Section 2 to the operational principles of safety filters.

2 Functional safety according to industrial standards and resulting challenges for data-driven methods

This section reviews established industrial standards on functional safety and outlines how they define and ensure functional safety. Moreover, it highlights the key challenges of applying these principles to data-driven methods in open environments and motivates the use of formal methods for ensuring functional safety.

2.1 Industrial standards on functional safety

Functional safety is the part of overall safety that depends on the correct functioning of a system and its control system, including all safety-related technologies. The definition and application of the functional safety concept are specified in international standards [16], which formalize the concept and prescribe structured processes across the safety life cycle. Industrial standards on functional safety specify objectives, structured processes, and work products across the safety life cycle to engineer and demonstrate functional safety [18], [26], [27]. They define how to identify hazards, derive and assign safety requirements, implement safety functions, and perform V&V with documented evidence throughout all development phases.

Across all industrial standards on functional safety, a central objective is to ensure that hazards are identified and risks are reduced to an acceptable level [16]. Classical safety standards [18], [26] define functional safety with respect to faults, whereas standards like [27] extend functional safety to the concept of Safety of the Intended Functionality (SOTIF), which addresses hazards arising from insufficient functionality, even in the absence of faults.

The processes defined in the functional safety life cycle are applied and refined in domain-specific standards, many of which build on IEC 61508 [18] as the central and most general standard for functional safety. In the automotive domain, ISO 26262 [26] and ISO 21448 [27] specify requirements for functional safety and SOTIF, while IEC 62061 [28] adapts the principles of IEC 61508 [18] to machinery. For a comprehensive overview of functional safety standards across specific application domains, we refer to [16].

At the regulatory level, the European Machinery Regulation (EU) 2023/1230 [29] establishes binding requirements for the design, construction, and conformity assessment of machinery. Moreover, the Commission Implementing Regulation (EU) 2022/1426 [17] defines type-approval procedures for automated driving systems. Both provide the legal framework for enabling harmonized safety standards.

Beyond domain-specific standards, emerging ISO standards and guidance documents address challenges posed by artificial intelligence (AI)-based[1] and data-driven methods. ISO/PAS 8800 [30] complements ISO 26262 [26] and ISO 21448 [27] by providing a structured framework to engineer road-vehicle systems that achieve and assure functional safety when incorporating AI-based technologies. Independently, ISO/IEC TR 5469 [25] examines the concept of functional safety for AI systems and provides guidance on their safe integration. Based on these key industrial standards, the fundamental concepts that define functional safety are presented in the next section.

2.2 General concept of functional safety

Functional safety standards [16] introduce the concept of harm as a fundamental basis for defining safety objectives and requirements:

Definition 1. Harm [18], [26]
Harm is physical injury or damage to the health of persons, property, or the environment.

Functional safety comprises the prevention or reduction of harm by requiring that safety functions are performed correctly and within the required time, addressing hazards[2] arising from malfunctioning behavior as well as from insufficient intended functionality or foreseeable misuse [18], [26], [27]. Thereby, all functional safety standards [16] rely on the concept of risk to describe the potential for harm:

Definition 2. Risk [16], [18], [26], [27]
Risk is the combination of the severity of potential harm and the probability that such harm occurs.

Within functional safety standards [16], the primary objective is to reduce risk to an acceptable level, and it is required that the reduction is demonstrated. Accordingly, it is a requirement that the control system (with or without

1 ISO/PAS 8800 [30] and ISO/IEC TR 5469 [25] use the term AI as defined in ISO/IEC 22989 [31]. Within the scope of this work, we refer to methods whose behavior is at least partially learned from data as data-driven methods, including approaches that may combine learned components with analytical or physics-based models.

2 A hazard is a potential source of harm [18], [26], [27].

data-driven methods) in an autonomous system facing a hazard must satisfy:[3]

Requirement 1. Risk Reduction

To ensure functional safety, risk must be demonstrably reduced to an acceptable level.

Risk reduction addresses the likelihood of hazardous behavior and the potential severity of the hazardous behavior. While different standards may formalize risk assessment differently, for example, by introducing a Safety Integrity Level (SIL)[4], the principle remains the same: safety measures must mitigate risk by reducing factors such as the occurrence rate of the hazardous behavior R_{HB}, the probability that hazardous behavior leads to a critical event $P_{\mathrm{E|HB}}$, the probability that the critical event cannot be controlled $P_{\mathrm{C|E}}$, and the probability that an uncontrollable event results in severe harm $P_{\mathrm{S|C}}$. Here, we emphasize the influence of the probability that the critical event cannot be controlled $P_{\mathrm{C|E}}$, which is directly influenced by the controllability of the system:

Definition 3. Controllability [26]

Controllability is the capability of a system to prevent hazardous behavior from causing harm, including the ability to detect critical conditions and transition to or maintain a safe condition.

The effectiveness of risk reduction can be quantified through an *acceptance criterion* A_{H}:

$$A_{\mathrm{H}} = P_{\mathrm{S|C}} \cdot P_{\mathrm{C|E}} \cdot P_{\mathrm{E|HB}} \cdot R_{\mathrm{HB}}. \tag{1}$$

Risk reduction aims to keep the risk below the acceptance criterion by mitigating the contributing factors like $P_{\mathrm{C|E}}$ or $P_{\mathrm{S|C}}$. The selection of an appropriate acceptance criterion A_{H} is typically guided by domain standards and regulatory requirements.

To ensure risk reduction to an acceptable level, functional safety standards require a structured development process [18], [26], [27] that systematically addresses hazards and ensures safe operation, as illustrated in Figure 1. At the core of this process is the assessment and control of risk, which distinguishes between a specification and design phase and an evaluation phase, where V&V demonstrate compliance with these requirements.

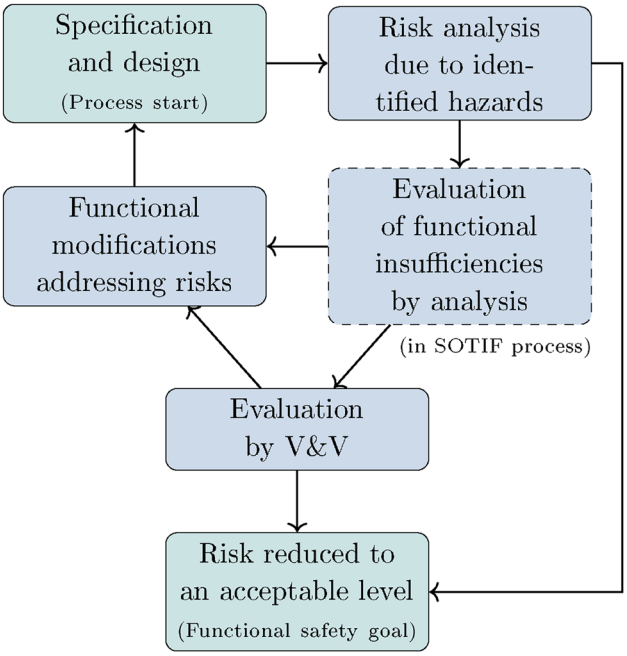

Figure 1: Simplified development cycle (based on ISO 21448 [27]) with iterative risk reduction.

The specification and design phase defines the system's intended functionality, performance requirements, mitigation measures, and system architecture. These elements are formalized as safety specifications:

Definition 4. Safety Specifications [27]

Safety specifications are documented descriptions of intended functionality, identified hazards, and associated mitigation measures within a system.

Within the development process, safety specifications must be established and maintained to ensure that hazards are systematically addressed [27]. The safety specifications can be refined iteratively with a hazard analysis and risk assessment (HARA) or by iteratively addressing hazards that arise from insufficient functionality according to SOTIF, even in the absence of faults [27]. Specifications are only formulated and valid within the system's operational boundaries, captured in the operational design domain (ODD):

Definition 5. ODD [20], [26], [27], [33]

The ODD is the formally specified set of conditions and constraints under which an automated system is intended to operate safely.

Beyond defining the ODD, the applicable safety standards [17], [26], [27] specify detection of when operation approaches or leaves this domain and initiation of a controlled fallback strategy. This strategy is intended to

3 Control-specific risk reduction is further detailed in ISO 13849 [32], which specifies safety requirements and performance levels for safety-related control systems in machinery.

4 While [18] introduces the SIL as a discrete measure of the required reliability of safety functions, [27] refines risk assessment by explicitly considering the factors *severity*, *exposure*, and *controllability*.

achieve a minimal risk condition, typically by means of a minimal risk manoeuvre, thereby supporting functional safety even outside the intended ODD.

In the evaluation phase of the development cycle, V&V activities demonstrate that the system reduces risk to an acceptable level by ensuring compliance with specifications across both known and unknown scenarios.

2.3 Challenges of ensuring functional safety in open environments

Systems operating in real-world environments are exposed to dynamic and unpredictable conditions, where not all relevant scenarios can be anticipated during specification and design, and complete knowledge of the environment is unattainable.

Definition 6. Open Environment

An open environment contains a set of scenarios, objects, and events, including those not anticipated during design, for which a complete specification is impractical at design time.

Unlike closed environments in which all possible scenarios can be specified in advance, such that safety can be ensured for all scenarios [20], open environments challenge the assumptions of traditional safety standards. The inherent variability of open environments makes it impossible to exhaustively specify, model, or test all relevant scenarios.

Challenge 1. Operation in an Open Environment

Open environments introduce inherent uncertainty into the scenarios an automated system may encounter.

Using detailed specifications is standard practice in IEC 61508 [18] and ISO 26262 [26]. However, for dynamic open environments, complete specifications are infeasible because either rare situations are practically unpredictable or the number of potential situations is too large for exhaustive enumeration [34]. Consequently, rare or unusual scenarios cannot be fully captured in the specifications.[5]

Challenge 2. Incomplete Specifications

It is impractical to specify all requirements to cover the enormous variety of events in an open environment.

Approaches like SOTIF [27] address Challenge 2 by iteratively refining the ODD and specifications, but this process remains time-consuming and cannot ensure that all relevant scenarios and triggering conditions are identified, in particular with respect to unknown scenarios as discussed in SOTIF [27]. As an additional challenge, brute-force validation of the method with detailed specifications might require impractical amounts of test time, particularly due to the long tail of rare events. For example, the authors in [35], [36] estimate that achieving an acceptance rate commonly required in the domain of automated driving would demand hundreds of millions of miles of testing.

Challenge 3. Testing Effort

Validating functional safety in open environments by testing might require an impractically large amount of testing.

While the concepts of an ODD and SOTIF provide a structured description of the intended operational conditions and thus partially address the functional safety challenges posed by open environments, the use of data-driven methods introduces additional difficulties for ensuring functional safety.

2.4 Challenges of ensuring functional safety using data-driven methods

Because large parts of the functionality of data-driven methods are encoded in the training data rather than specified explicitly, the safety argument of such methods critically depends on the representativeness and the sufficiency of the data, both of which are inherently difficult to guarantee. Edge-case data, including conditions at the boundaries of the ODD, are rare and may be safety-critical. Ensuring their representative inclusion in the training data of data-driven methods is costly and time-consuming, so the data is likely to remain incomplete, which might result in insufficient safety coverage [20], [37].

Challenge 4. Insufficient Coverage

The rarity and cost of collecting edge-case data result in incomplete coverage of safety-relevant functionality in data-driven methods.

Even if the data captured at a single point in time were sufficient to cover the safety-relevant functionality of the

5 For example, in road traffic, the combinatorial variety of interacting elements, such as unexpected pedestrian movements, erratic behavior of other drivers, sudden weather changes, or infrastructure faults, results in a set of possible situations for which a fully exhaustive specification is infeasible in advance [19].

method, real-world conditions may evolve beyond what is represented in the training data [38].

Challenge 5. Data drift

In open environments, changes over time can cause data-driven systems to encounter unseen situations.

Furthermore, data-driven methods utilizing complex models such as deep neural networks are difficult to interpret, and their validation requires dedicated validation methods [20].

Challenge 6. Lack of Interpretability

Data-driven models often lack explanations that are understandable to humans, limiting the transparency of their internal decision logic.

The challenges arising from the use of data-driven methods (see Challenges 4–6) lead to an even more extensive safety development process to ensure their safe integration. ISO/PAS 8800 [30] complements ISO 26262 [26] and ISO 21448 [27] by embedding these additional assurance demands into established functional safety processes, ensuring that the resulting requirements are systematically captured and addressed. Nevertheless, this does not reduce the fundamental testing effort but rather amplifies it, which is why ISO/PAS 8800 [30] and related standards also highlight the need for mathematically supported techniques to strengthen controllability and overall safety assurance.

2.5 Towards functional safety for data-driven methods by means of formal methods

In contrast to data-driven methods and their associated challenges (see Challenges 4–6), formal methods are examined as a means of providing mathematically rigorous specifications and proofs in many industrial standards. IEC 61508 [18] explicitly recommends the use of formal methods for the specification, design, and verification of safety-related systems, particularly for applications with higher SILs. Formal methods provide mathematically based techniques that enable precise and unambiguous system models, supporting the completeness, consistency, and correctness of specifications and implementations. By enabling safety arguments through mathematical proof, formal methods might reduce reliance on extensive empirical validation (see Challenge 3) and help avoid impractical testing scenarios as motivated by [20]. Complementary formal and analytical techniques are therefore encouraged to provide the necessary safety evidence beyond what empirical testing alone can deliver.

To enable the application of formal methods to data-driven approaches, ISO 21448 [27] and ISO/IEC TR 5469 [25] emphasize the importance of *monitoring and intervention* as runtime mechanisms for maintaining functional safety. ISO/PAS 8800 [30] reinforces this view by requiring supervisory and limiting logic, potentially including non-AI backup functions, to constrain AI-based behavior within predefined safety limits and to ensure controllability when errors in the AI-based method are detected, for example due to unexpected inputs or model insufficiencies.

ISO/IEC TR 5469 [25] highlights mathematically grounded approaches as promising means to formally implement intervention strategies and to define provably safe operational bounds that can be enforced at runtime. These approaches ensure that, even if data-driven components have errors or execute incorrect control actions, the overall system continues to operate within the safe bounds. Figure 2 depicts such a monitoring and intervention strategy. Continuous monitoring verifies that the operating conditions of the data-driven components remain within the assumptions made during development. If deviations occur, the system must transition to a safe condition.[6] To make this feasible, the intervention provides an additional protection layer by switching to a backup decision system.

Recommendation 1. Monitoring and Intervention

Data-driven methods should be continuously monitored to detect deviations from assumed conditions. When deviations occur, appropriate interventions should ensure a transition to a safe condition.[7]

3 Formalizing safety for autonomous systems

Building on the functional safety concepts introduced in Section 2, this section establishes the formal framework used throughout the paper. We discuss safety paradigms aligned with functional safety and formalize the notion of safety, with particular emphasis on all-time safety. We begin by introducing a general model of a system operating in an open environment.

Consider the dynamics of a system to be given by the differential equation

6 As defined in ISO 26262 [26] and ISO 21448 [27], i.e., conditions of the functional state without unreasonable risk.

7 ISO 21448 [27], ISO/IEC TR 5469 [25] and ISO/PAS 8800 [30] highlight continuous monitoring and non-AI backup or supervisory functions as essential measures to maintain functional safety when AI- or data-driven methods encounter deviations from their intended operating conditions.

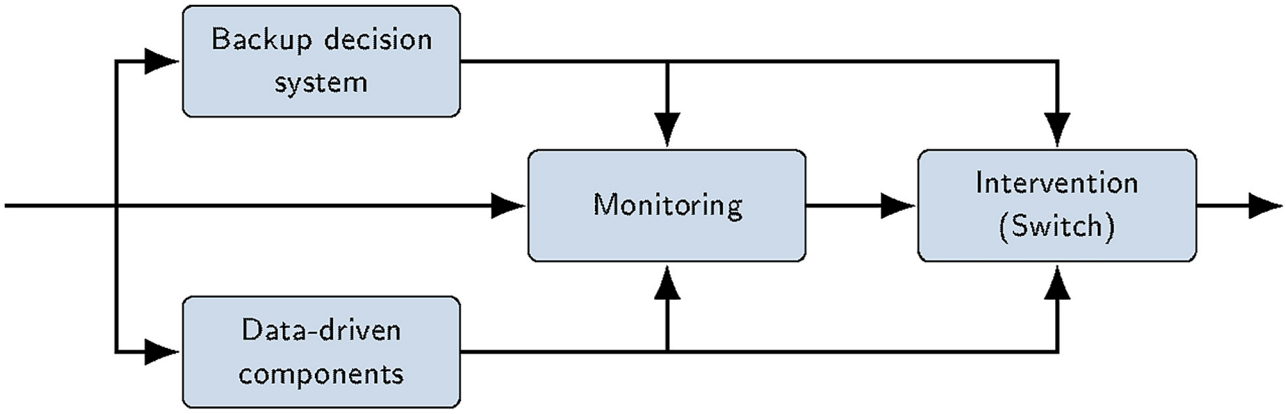

Figure 2: Schematic architecture of a monitoring and intervention strategy (based on ISO/IEC TR 5469 [25]). The available information is provided to the data-driven components, the monitoring component, and the backup decision system. The data-driven components generate the nominal output, while the backup decision system computes a backup output. The monitoring component checks whether the predefined safety conditions are satisfied. Based on the outcome of the monitoring component, the intervention component determines which output is applied.

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{d}(t)), \tag{2}$$

with the system state $\boldsymbol{x}(t) \in \mathcal{X} \subset \mathbb{R}^{n_x}$, the control input $\boldsymbol{u}(t) \in \mathcal{U} \subset \mathbb{R}^{n_u}$, the disturbance $\boldsymbol{d}(t) \in \mathbb{R}^{n_d}$, and $\boldsymbol{f} : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_d} \to \mathbb{R}^{n_x}$. The disturbance $\boldsymbol{d}(t)$ models external disturbances (e.g., wind) and predictive uncertainty in the system dynamics, including model mismatch (e.g., due to uncertain physical parameters). The sets $\mathcal{X}$ and $\mathcal{U}$ reflect physical limitations as well as design constraints of the system (e.g., maximum actuator force or maximum velocity). In addition, the system (2) is subject to environment-induced constraints on $\boldsymbol{x}(t)$, such as obstacles. The existence of such environmental constraints constitutes a hazard,[8] since collision with an obstacle may lead to harm (see Definition 1). For a given hazardous event in which environmental constraints exist, we collect all states corresponding to harm into the set of *harmful states* $\mathcal{H} \subset \mathbb{R}^{n_x}$. In the presence of hazardous events, the system state $\boldsymbol{x}(t)$ must avoid the set of harmful states $\mathcal{H}$. In our notation, this requirement is expressed by constraining the system state $\boldsymbol{x}(t)$ to remain within the set of *constraint-satisfying states* $\mathcal{X}_{\mathcal{H}} \subset \mathbb{R}^{n_x}$, with $\mathcal{X}_{\mathcal{H}} := \mathcal{X} \backslash \mathcal{H}$.

3.1 Safety paradigms

The safety of system (2) within $\mathcal{X}_{\mathcal{H}}$ must be ensured, in particular in the presence of disturbances $\boldsymbol{d}(t)$. Depending on how disturbances are modeled, different safety paradigms and corresponding safety definitions arise. If the disturbance $\boldsymbol{d}(t)$ is modeled as a stochastic process with an associated probability distribution, probabilistic notions of safety can be employed. In this case, safety is typically defined in a *risk-aware* paradigm by bounding the tail risk of a trajectory-dependent cost or constraint-violation measure. A comprehensive overview of risk-aware safety concepts is provided in [39].

Alternatively, disturbances can be modeled as bounded but otherwise arbitrary, such that $\boldsymbol{d}(t) \in \mathcal{D}$, with the bounded set $\mathcal{D} \subset \mathbb{R}^{n_d}$. This leads to a robust *worst-case* paradigm of safety. Under this paradigm, the system is considered safe if it never enters the set of harmful states $\mathcal{H}$, even under the worst-case disturbance $\boldsymbol{d}(t) \in \mathcal{D}$.

Both paradigms can be related to the notion of functional safety (see Section 2.2), which requires safety to be demonstrated within a predefined ODD. For instance, an automated vehicle may be required to operate safely under crosswinds. In the risk-aware paradigm, disturbances such as the wind speed are modeled as a stochastic process, and it must be demonstrated that the resulting probability of harm remains below the acceptance criterion A_{H}. In the worst-case paradigm, safety is established by showing that any disturbance realization within the ODD (e.g., a specified maximum wind speed) does not lead to harm. While disturbances in the ODD may induce hazardous behavior, harm can be avoided as long as the resulting critical events remain controllable. Disturbances beyond the specification are either excluded from the analysis as operation outside the ODD or shown to occur with sufficiently low probability, that is, $R_{\mathrm{HB}} \approx 0$ in (1) within the HARA framework.

8 In an open environment, the system may face infinitely many scenarios with potential hazards (see Definition 6). Unlike physical constraints $\mathcal{X}$, environmental constraints on $\boldsymbol{x}(t)$ are generally unknown *a priori* and only perceived during operation.

3.2 Automation design

A general automation concept for the system (2) comprises four core functional automation modules: localization, perception, planning, and control [40]. Localization provides an estimate of the system state $\boldsymbol{x}(t)$, while perception enables sensing and interpretation of the environment, which also includes the identification of harmful states $\mathcal{H}$. Based on this information, the planning module generates a desired behavior, and the control module computes a corresponding control input $\boldsymbol{u}(t)$. In practical implementations, all modules influence whether applying a control input $\boldsymbol{u}(t)$ is safe with respect to the current state $\boldsymbol{x}(t)$ and the harmful states $\mathcal{H}$. Since each module is subject to uncertainty, these uncertainties must be explicitly considered in the design process in accordance with functional safety principles. Localization, which is typically realized through state estimation, is affected by measurement noise and modeling errors. Methods for computing state-estimation error bounds are discussed in [41]. Furthermore, a discussion on safety filters that explicitly account for imperfect perception is provided in [21]. In this work, however, we assume that the system state $\boldsymbol{x}(t)$ is perfectly measurable in order to isolate the effect of disturbances $\boldsymbol{d}(t)$ within the system dynamics (2). These disturbances therefore constitute the primary source of uncertainty affecting the functional safety of the control module.

The control module for system (2) is realized via a state-feedback law $\boldsymbol{u}(t) = \boldsymbol{\pi}(\boldsymbol{x}(t))$, where $\boldsymbol{\pi}\colon \mathbb{R}^{n_x} \to \mathbb{R}^{n_u}$ denotes the feedback control strategy. This yields the autonomous closed-loop system

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{\pi}(\boldsymbol{x}(t)), \boldsymbol{d}(t)). \tag{3}$$

In this work, we focus on analyzing whether a control strategy $\boldsymbol{\pi}$ can be designed to ensure safety for the closed-loop system (3), assuming exact knowledge of both the harmful states $\mathcal{H}$ and the system state $\boldsymbol{x}(t)$.

3.3 All-time safety

In the following, we restrict attention to the worst-case safety paradigm, in which the disturbance $\boldsymbol{d}(t)$ is assumed to be bounded but otherwise arbitrary within $\mathcal{D}$. However, simply demonstrating that the system state $\boldsymbol{x}(t)$ lies within $\mathcal{X}_{\mathcal{H}}$ at a given time is not sufficient to ensure safety. Due to input constraints and disturbances, the system (2) may still inevitably leave $\mathcal{X}_{\mathcal{H}}$ in the future, regardless of the applied control input $\boldsymbol{u}(t)$. States from which a collision with $\mathcal{H}$ is inevitable under an unfavorable disturbance realization, or which are inherently unsafe, are referred to as *potential collision states*.

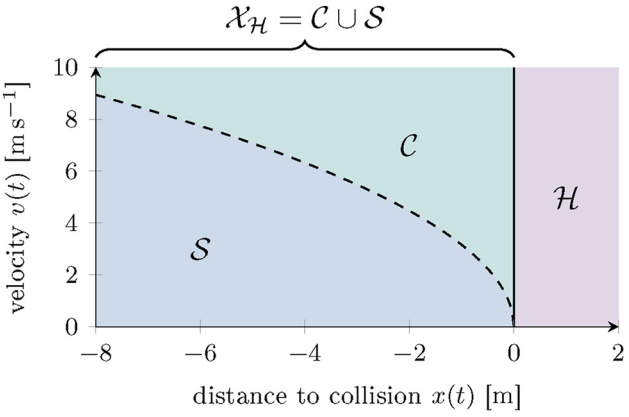

Figure 3: Partition of the constraint-satisfying states $\mathcal{X}_{\mathcal{H}}$ into potential collision states $\mathcal{C}$ and safe states $\mathcal{S}$ for the system $\dot{x}(t) = v(t),\ \dot{v}(t) = a(t),\ |a(t)| \leq a_{\max}$ with maximal braking force $a_{\max} = 5$ m s^{-2} approaching $\mathcal{H} = \{(\bar{x}, \bar{v}) \in \mathbb{R}^2 \mid \bar{x} > 0\}$.

Definition 7. Potential Collision States

The set of potential collision states $\mathcal{C}$ consists of all states $\boldsymbol{x}(0) \in \mathcal{X}_{\mathcal{H}}$ such that for (2) it holds:

$$\exists \boldsymbol{d},\ \forall \boldsymbol{u},\ \exists t > 0\colon\ \boldsymbol{x}(t) \in \mathcal{H}, \tag{4}$$

with $\boldsymbol{u}\colon \mathbb{R}_+ \to \mathcal{U}$ and $\boldsymbol{d}\colon \mathbb{R}_+ \to \mathcal{D}$.

Hence, the constraint-satisfying states $\mathcal{X}_{\mathcal{H}}$ are partitioned into the potential collision states $\mathcal{C}$ and *safe states* $\mathcal{S} \subset \mathbb{R}^{n_x}$ with $\mathcal{S} := \mathcal{X}_{\mathcal{H}} \backslash \mathcal{C}$, as illustrated in Figure 3.

For safe operation in open environments, a feedback strategy $\boldsymbol{\pi}$ must explicitly account for the harmful states $\mathcal{H}$. Safety of the closed-loop system (3) is therefore defined with respect to the existence of safe states $\mathcal{S}$. In particular, safety requires that the feedback strategy $\boldsymbol{\pi}$ ensures *all-time safety*.

Definition 8. All-Time Safety

The autonomous system (3) is all-time safe if the feedback strategy $\boldsymbol{\pi}$ ensures that

$$\boldsymbol{x}(t) \in \mathcal{S} \quad \text{for all } t \geq 0. \tag{5}$$

Hence, to ensure safety of the autonomous system (3), it must be ensured that the state $\boldsymbol{x}(t)$ never leaves the set of safe states $\mathcal{S}$. To this end, safety filters are outlined in the following as a method to ensure the property of all-time safety.

4 Safety filter

We describe the switching safety filter as the general concept of safety filters in Section 4.1, and present design methods in Section 4.2, as well as approaches for the computation of safe sets in Section 4.3. Optimization-based runtime implementations are discussed in Section 4.4. Section 4.5 demonstrates the design of different safety filter types in a simple collision-avoidance example.

4.1 General concept of safety filters

A safety filter is an intervention scheme that ensures all-time safety (see Definition 8) for an autonomous system (3) when operated under a nominal control strategy, even if the nominal control strategy alone cannot ensure safety [21]. To address the challenges outlined in Section 2.3, we use a data-driven feedback law $\boldsymbol{u}^{\mathrm{dd}}(t) = \boldsymbol{\pi}^{\mathrm{dd}}(\boldsymbol{x}(t))$, with $\boldsymbol{\pi}^{\mathrm{dd}} : \mathbb{R}^{n_x} \to \mathbb{R}^{n_u}$ as the nominal control strategy.

Safety filters rely on the existence of a feedback strategy $\boldsymbol{\pi}^{\Omega} : \mathcal{X} \to \mathcal{U}$ that ensures *positive set invariance*. To ensure safety according to the worst-case paradigm, positive set invariance also requires a robust definition with a bounded worst-case disturbance in $\mathcal{D}$.

Definition 9. Robust Invariant Set

A set $\Omega \subseteq \mathbb{R}^{n_x}$ is robustly invariant for (3) with $\boldsymbol{\pi} = \boldsymbol{\pi}^{\Omega}$, if for any state

$$\boldsymbol{x}(0) \in \Omega \;\Rightarrow\; \forall t > 0,\ \forall \boldsymbol{d} :\ \boldsymbol{x}(t) \in \Omega. \tag{6}$$

Definition 10. Safe Set

A safe set is a robustly invariant set Ω for which $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ holds. The largest safe set coincides with the set of safe states $\mathcal{S}$. Therefore, any safe set can equivalently be described as a subset of the safe states, i.e., $\Omega \subseteq \mathcal{S}$.

Based on this concept of set invariance, the safety filter is formulated. The simplest realization of a safety filter uses a safe set $\Omega \subseteq \mathcal{S}$ and switches to $\boldsymbol{\pi}^{\Omega}$ on the boundary of Ω ($\partial\Omega$) as a *backup control* strategy.

Definition 11. Safety Filter

A safety filter $\boldsymbol{\pi}^{\mathrm{safe}} : \mathcal{X}_{\mathcal{H}} \times \mathbb{R}^{n_u} \to \mathcal{U}$ supervises the data-driven feedback law $\boldsymbol{u}^{\mathrm{dd}}(t) = \boldsymbol{\pi}^{\mathrm{dd}}(\boldsymbol{x}(t))$ with the switching intervention scheme:

$$\boldsymbol{\pi}^{\mathrm{safe}}(\boldsymbol{x}(t), \boldsymbol{u}^{\mathrm{dd}}(t)) := \begin{cases} \boldsymbol{\pi}^{\Omega}(\boldsymbol{x}(t)), & \boldsymbol{x}(t) \in \partial\Omega, \\ \boldsymbol{u}^{\mathrm{dd}}(t), & \text{otherwise}, \end{cases} \tag{7}$$

with the safe set $\Omega \subseteq \mathcal{S}$, for which robust invariance is ensured by the backup controller $\boldsymbol{\pi}^{\Omega}(\boldsymbol{x}(t))$.

The safety filter is composed of distinct modules, as depicted in Figure 4. The first module is a *monitor*, which observes the system state $\boldsymbol{x}(t)$. If the monitor detects that the system is leaving Ω at $\partial\Omega$, the safety filter intervenes with an *intervention* strategy by switching to the backup control law $\boldsymbol{u}(t) = \boldsymbol{\pi}^{\Omega}(\boldsymbol{x}(t))$.

4.2 Safety filter design

To make safety filters applicable, it is essential to translate the general principle of safety filters in Definition 11 into design methods. Comprehensive overviews of design methods are provided in [22], [42]. Two key challenges in current research critically influence the applicability of safety filters:

1. The degree of freedom available to the data-driven controller $\boldsymbol{\pi}^{\mathrm{dd}}$ strongly depends on the size of the used invariant set Ω for the safety filter. The larger the set Ω is, the less frequently the filter needs to switch to the backup controller $\boldsymbol{\pi}^{\Omega}$. Ideally, Ω should coincide with

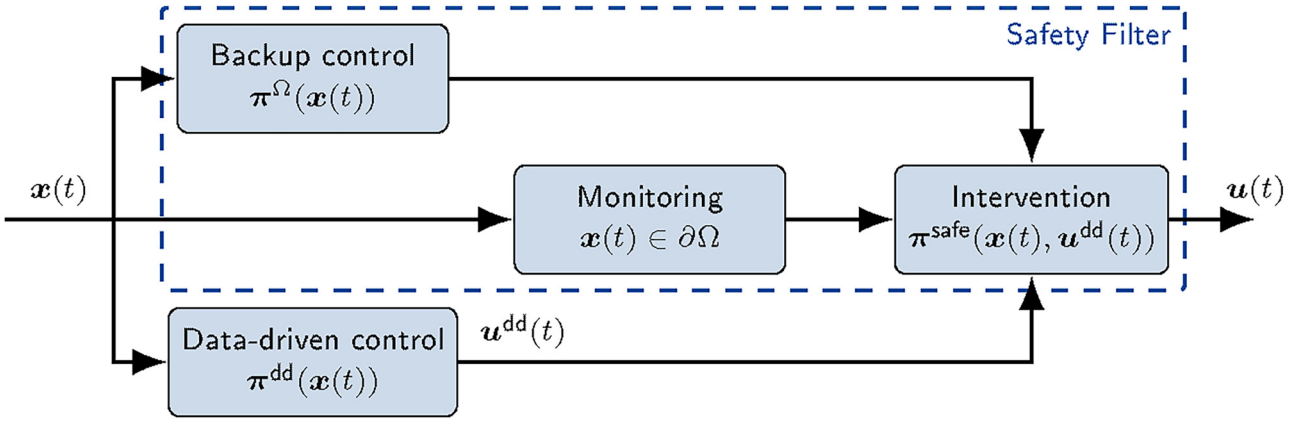

Figure 4: Schematic architecture of the switching safety filter introduced in Definition 11 (based on [21]).

the safe states $\mathcal{S}$, as this minimizes interventions. This motivates the notion of a *least restrictive safety filter*. However, computing or approximating $\mathcal{S}$ is challenging for nonlinear systems with input constraints and disturbances, which constitutes a central problem in safety filter design.

2. Regardless of the size of the safe set, challenges arise when the data-driven controller $\boldsymbol{\pi}^{\mathrm{dd}}$ generates unsafe inputs that drive the state onto the boundary $\partial\Omega$. Then, the safety filter (11) switches to $\boldsymbol{\pi}^{\Omega}$ on the boundary $\partial\Omega$ of the safe set to recover the state in the interior of Ω, after which $\boldsymbol{\pi}^{\mathrm{dd}}$ may again drive the state onto the boundary. This frequent switching between $\boldsymbol{\pi}^{\Omega}$ and $\boldsymbol{\pi}^{\mathrm{dd}}$ can lead to chattering.

In the following, we focus on the open practical challenges that determine the applicability of safety filters as risk-reduction measures in the context of functional safety. The approach used to compute invariant sets directly determines whether a safety filter is practically applicable. For this reason, Section 4.3 focuses on applicable methods to compute a control invariant set $\Omega \subseteq \mathcal{S}$. Since chattering can restrict real-world deployment, Section 4.4 examines alternative intervention strategies to the switching safety filter.

4.3 Computation of safe sets

Determining a feasible safe set $\Omega \subseteq \mathcal{S}$ in a dynamic environment, where the unsafe region $\mathcal{H}$ changes over time, is challenging. The set must simultaneously satisfy being in $\mathcal{X}_{\mathcal{H}}$, and set invariance must be ensured with a corresponding feedback strategy $\boldsymbol{\pi}^{\Omega}$.

While there exist methods for online computation of such sets [43], [44], they typically rely on convexification of set calculations and linearization of the dynamics of the system (2). Therefore, these methods ensure safety only for under-approximations of the safe states $\Omega \subseteq \mathcal{S}$, which makes the methods more conservative. A common alternative is to determine and validate the invariance of a candidate set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ offline, where the constraint-satisfying states $\mathcal{X}_{\mathcal{H}}$ denote a predefined region outside of which system operation is considered potentially harmful, and to enforce the condition $\mathcal{X}_{\mathcal{H}} \cap \mathcal{H} = \emptyset$ during online application [45], [46].

The mathematical foundation for computing invariant sets $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ is *viability theory* [47], which studies the evolution of constrained dynamical systems. A fundamental result is *Nagumo's theorem* [48].

Definition 12. Nagumo's Theorem

Let $h\colon \mathbb{R}^{n_x} \to \mathbb{R}$ be a continuously differentiable scalar function defining the set Ω as

$$\begin{aligned} \Omega &:= \{\bar{\boldsymbol{x}} \in \mathbb{R}^{n_x} : h(\bar{\boldsymbol{x}}) \geq 0\}, \\ \partial\Omega &:= \{\bar{\boldsymbol{x}} \in \mathbb{R}^{n_x} : h(\bar{\boldsymbol{x}}) = 0\}, \end{aligned} \tag{8}$$

then, Ω is forward invariant for an autonomous system $\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t))$ if

$$\dot{h}(\boldsymbol{x}(t)) = \nabla h(\boldsymbol{x}(t))\boldsymbol{f}(\boldsymbol{x}(t)) \geq 0, \quad \forall \boldsymbol{x}(t) \in \partial\Omega. \tag{9}$$

This condition ensures that the value of h does not decrease on $\partial\Omega$, and therefore trajectories have to remain in Ω. Nagumo's theorem provides necessary and sufficient conditions for the invariance of a closed set. A further central concept in viability theory is the *viability kernel*, defined as the set of initial conditions from which constraint satisfaction can be maintained indefinitely. Determining the safe states $\mathcal{S}$ for (2) corresponds to determining the viability kernel of $\mathcal{X}_{\mathcal{H}}$ under admissible inputs $\boldsymbol{u}(t) \in \mathcal{U}$ and, in the robust case, against bounded disturbances $\boldsymbol{d}(t) \in \mathcal{D}$. This robust case naturally leads to formulations of viability theory in terms of differential games. In this setting, with a worst-case disturbance $\boldsymbol{d}(t)$, the robustly invariant set (see Definition 9) is described by the outcome of a two-player zero-sum differential game: one player, selecting $\boldsymbol{u}(t)$, aims to keep (2) away from harmful states $\mathcal{H}$, while the opposing player, selecting $\boldsymbol{d}(t)$, aims to drive (2) into $\mathcal{H}$ [49].

A rigorous approach for computing the maximal safe set $\mathcal{S}$ for such a game is to leverage Hamilton-Jacobi (HJ) reachability analysis [50]. In HJ reachability analysis, the differential game is solved numerically via dynamic programming. This introduces numerical errors [46], [51] and suffers from the curse of dimensionality, since computational cost grows exponentially with system dimension [52]. To alleviate these limitations, methods such as system decomposition [53], warm-starting [54], and adaptive grids [55] are proposed.

Despite these computational challenges and numerical errors, HJ reachability analysis remains highly attractive, as it yields the maximal safe set $\mathcal{S}$, and therefore enables computing least-restrictive filters. Using the maximal safe set $\mathcal{S}$, the safety filter as formulated in Definition 11 can be instantiated with $\Omega = \mathcal{S}$, yielding a least restrictive switching safety filter. The practical applicability of such least restrictive safety filters has been demonstrated across a range of applications, including quadrotors [56], safe robot navigation [57], motion planning [45], [58], and flight envelope protection [59].

One of the most widely used approaches for constructing invariant sets for the autonomous system (3) builds on Nagumo's theorem through control barrier functions (CBFs) [24], [60]–[62]. Analogously to Lyapunov functions for stability, CBFs serve as safety certificates: rather than certifying convergence to an equilibrium, they certify invariance of Ω [24]. In the standard formulation, CBFs are formulated for nominal system dynamics, i.e., assuming the absence of external disturbances ($\boldsymbol{d}(t) = \mathbf{0}$). The function h in (8) is a valid CBF for Ω if there exists an extended class-$\mathcal{K}$ function[9] α such that, for all $\bar{\boldsymbol{x}} \in \Omega$,

$$\sup_{\bar{\boldsymbol{u}} \in \mathcal{U}} \nabla h(\bar{\boldsymbol{x}})\, \boldsymbol{f}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \mathbf{0}) \geq -\alpha(h(\bar{\boldsymbol{x}})) \tag{10}$$

holds. For a state $\bar{\boldsymbol{x}} \in \partial\Omega$, the CBF reduces to $h(\bar{\boldsymbol{x}}) = 0$, and the inequality (10) reduces to $\sup_{\bar{\boldsymbol{u}} \in \mathcal{U}} \nabla h(\bar{\boldsymbol{x}})\, \boldsymbol{f}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \mathbf{0}) \geq 0$, thereby ensuring Nagumo's invariance condition (9). Beyond this boundary case, CBFs extend the concept of Nagumo's theorem to the interior of the safe set via a continuous inequality constraint. This allows safety requirements to be enforced proactively before the system reaches $\partial\Omega$, yielding smoother behavior. In general, multiple control inputs may satisfy (10). In this formulation, the safety filter can select a control input from the admissible set $\{\bar{\boldsymbol{u}} \in \mathcal{U} \mid \nabla h(\bar{\boldsymbol{x}})\, \boldsymbol{f}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \mathbf{0}) \geq -\alpha(h(\bar{\boldsymbol{x}}))\}$, which characterizes all inputs satisfying the CBF constraint. In practice, an admissible input closest to the nominal control $\boldsymbol{u}^{\mathrm{dd}}(t)$ is computed online by solving an optimization problem subject to (10) (see Section 4.4).

Most CBF design methods presented in the literature [24], [60], [62] are tailored to specific system classes. In contrast, sum-of-squares (SOS) programming has found widespread use [63]. SOS methods have been successfully applied in safety-critical control problems such as robust motion planning via funnel computation [64], and automotive lane keeping and adaptive cruise control [24], [65]. In this framework, the barrier function h is parameterized as a polynomial of fixed degree, and SOS optimization [66] is employed to enforce the required CBF conditions. However, SOS-based approaches typically require the system dynamics to be control-affine and polynomial, which significantly limits their applicability to real-world systems with non-polynomial nonlinearities or non-affine control inputs.

These limitations reflect a more general challenge of CBF-based methods: constructing valid CBFs is often non-trivial, particularly for higher-order dynamics or in the presence of input constraints [62], where analytical constructions of h covering $\mathcal{S}$ rarely exist. Moreover, hand-crafted CBFs are typically conservative and rarely approximate the maximal safe set $\mathcal{S}$. To mitigate this, recent work seeks to refine pre-certified CBFs using HJ reachability [67] or to directly construct CBF-like functions from specialized HJ formulations [68] that recover $\mathcal{S}$ and yield a similar smooth decay of h as CBFs when approaching $\partial\Omega$. However, these approaches inherit the computational limitations and numerical errors of HJ reachability.

In their standard formulation (10) without external disturbances, CBFs provide limited robustness guarantees. As a result, several extensions explicitly account for model uncertainty and disturbances. These include robust CBFs [69]–[73], which enforce robust invariance for Ω if for all $\bar{\boldsymbol{x}} \in \Omega$ the worst-case formulation

$$\sup_{\bar{\boldsymbol{u}} \in \mathcal{U}} \inf_{\bar{\boldsymbol{d}} \in D} \nabla h(\bar{\boldsymbol{x}})\, \boldsymbol{f}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}, \bar{\boldsymbol{d}}) \geq -\alpha(h(\bar{\boldsymbol{x}})) \tag{11}$$

holds. Alternative approaches include input-to-state safe CBFs [74], [75], which provide robustness margins with respect to bounded disturbances, as well as probabilistic and stochastic CBFs [76]–[78], which incorporate uncertainty through chance constraints or stochastic dynamics models. Also, recent research [79], [80] investigates the construction of model-free CBFs, thereby eliminating the reliance on explicit system models and avoiding the need to explicitly account for model uncertainties.

4.4 Safety filter intervention based on online optimization

Directly applying the switching safety filter (see Definition 11) to activate the backup controller may induce chattering, as discussed in Section 4.2. Therefore, the backup controller $\boldsymbol{\pi}^{\Omega}$ is typically applied for at least one sampling interval when applied in discrete time instances, and often for longer durations, by design. Short activation intervals increase chattering, whereas longer intervals reduce reliance on the nominal strategy $\boldsymbol{\pi}^{\mathrm{dd}}$.

To overcome the switching behavior, many practical implementations replace the switching in Definition 11 by continuously modifying the nominal control input via online optimization. Following the switching safety filter as a baseline, we now discuss two major classes of optimization-based runtime implementations: CBF-based filters and predictive safety filters.

The CBF-based filters employ a continuously differentiable function $h\colon \mathbb{R}^{n_x} \to \mathbb{R}$ as a CBF satisfying (10). The safety filter computes a minimally invasive control input by

9 An extended class-$\mathcal{K}$ function is a continuous, strictly increasing function $\alpha\colon (-c, d) \to \mathbb{R}$ (with $c, d > 0$) that satisfies $\alpha(0) = 0$.

solving, at each time step, the optimization problem

$$\begin{aligned}\boldsymbol{\pi}^{\text{safe}}(\boldsymbol{x}(t), \boldsymbol{u}^{\text{dd}}(t)) := \arg\min_{\bar{\boldsymbol{u}}\in\mathbb{R}^{n_u}} \ & \frac{1}{2}\|(\bar{\boldsymbol{u}} - \boldsymbol{u}^{\text{dd}}(t))\|^2 \\ \text{s.t. } & \nabla h(\boldsymbol{x}(t))\, \boldsymbol{f}(\boldsymbol{x}(t), \bar{\boldsymbol{u}}, \boldsymbol{0}) \geq -\alpha(h(\boldsymbol{x}(t))), \\ & \bar{\boldsymbol{u}} \in \mathcal{U}.\end{aligned} \tag{12}$$

If (12) constitutes a feasible optimization problem, its solution yields the closest admissible input to the nominal control $\boldsymbol{u}^{\text{dd}}(t)$ while enforcing forward invariance of the safe set [24].

The CBF constraint activates smoothly as the system state $\boldsymbol{x}(t)$ approaches the boundary $\partial\Omega$ while ensuring that the system state $\boldsymbol{x}(t)$ remains inside Ω. This ensures safety while keeping the applied input close to the nominal one. As a result, CBF-based filters can be interpreted as smooth variants of switching safety filters, avoiding the chattering associated with hard switching. If the system (2) is control-affine, problem (12) becomes a quadratic program (QP) that can be solved efficiently online.

A further class of methods that follows an alternative intervention strategy compared to the switching approach in Definition 11 is given by model predictive safety filters (MPSFs). While safety filters based on CBFs enforce safety through local differential constraints, MPSF employs model predictive techniques similar to classical model predictive control (MPC) and imposes a safe set as a terminal constraint [81]. The central motivation is that explicitly computed invariant sets Ω are often conservative or computationally intractable, as previously discussed. Instead, MPSF can employ such a conservative safe set $\Omega^{\text{trm}} \subseteq \mathcal{S}$ and implicitly enlarge the admissible operating region through online trajectory optimization. However, during runtime, the state $\boldsymbol{x}(t)$ is not required to lie inside the terminal set Ω^{trm}. Instead, it suffices that the system can be driven into Ω^{trm} within a finite prediction horizon $T \in \mathbb{R}_{>0}$, while satisfying the state and input constraints $\mathcal{X}_{\mathcal{H}}$ and $\mathcal{U}$. This requirement is enforced through the following optimization problem:

$$\begin{aligned}\boldsymbol{u}^*(\cdot) := \arg\min_{\boldsymbol{u}(\cdot)} \ & \frac{1}{2}\|\boldsymbol{u}(0) - \boldsymbol{u}^{\text{dd}}(t)\|^2 \\ \text{s.t. } & \boldsymbol{x}(0) = \boldsymbol{x}(t), \\ & \dot{\boldsymbol{x}}(\tau) = \boldsymbol{f}(\boldsymbol{x}(\tau), \boldsymbol{u}(\tau), \boldsymbol{0}),\ \forall \tau \in [0, T], \\ & \boldsymbol{x}(\tau) \in \mathcal{X}_{\mathcal{H}},\ \forall \tau \in (0, T], \\ & \boldsymbol{u}(\tau) \in \mathcal{U},\ \forall \tau \in (0, T], \\ & \boldsymbol{x}(T) \in \Omega^{\text{trm}}.\end{aligned} \tag{13}$$

When applying the control strategy as a safety filter, the nominal input $\boldsymbol{u}^{\text{dd}}(t)$ is applied if it allows a trajectory to satisfy the constraints in (13). Otherwise, it is modified in a minimal manner according to the quadratic objective. The applied input is then given by

$$\boldsymbol{\pi}^{\text{safe}}(\boldsymbol{x}(t), \boldsymbol{u}^{\text{dd}}(t)) = \boldsymbol{u}^*(0). \tag{14}$$

The set of all states from which a predicted trajectory into Ω^{trm} exists implicitly defines a virtual safe set $\Omega' \subseteq \mathcal{S}$ with $\Omega^{\text{trm}} \subset \Omega'$. In this sense, the MPSF extends the admissible operating region beyond the conservative terminal set Ω^{trm}. As $T \to \infty$, the condition in (13) recovers the conditions for all-time safety in Definition 8, and the virtual safe set Ω' converges to the maximal safe set $\mathcal{S}$.

In practice, (13) is solved online by discretizing the system dynamics and applying a shooting method, resulting in a standard MPC problem. For nonlinear system dynamics (2), in contrast to CBF-QPs, the MPSF optimization problem is a dynamic optimization problem, for which worst-case computation time and solver reliability become a challenge.

Several extensions have been proposed to account for disturbances and uncertainty, as well as to establish additional stability properties [82]–[86]. Comprehensive overviews and unifying perspectives on runtime safety filter architectures, including barrier-based, reachability-based, and predictive approaches, are provided in [21], [22].

4.5 Collision-avoidance example

We consider the setup initially depicted in Figure 3 of a collision-avoidance scenario. The dynamics are modeled as a double integrator

$$\dot{x}(t) = v(t), \qquad \dot{v}(t) = a(t), \tag{15}$$

with the control input $a(t) \in \mathbb{R}$ representing the braking acceleration subject to the constraint $|a(t)| \leq a_{\max}$. The maximal braking capability is set to $a_{\max} = 5\ \text{m s}^{-2}$. The set of harmful states is given by $\mathcal{H} = \{(\bar{x}, \bar{v}) \in \mathbb{R}^2 \mid \bar{x} > 0\}$, corresponding to a collision.

For this simple system, the maximal safe set can be derived analytically as

$$\mathcal{S} := \left\{ (\bar{x}, \bar{v}) \in \mathbb{R}^2 \middle| \ \bar{x} \leq -\frac{\bar{v}^2}{2a_{\max}} \right\}, \tag{16}$$

which characterizes all states from which the system can still be brought to rest before reaching the harmful states $\mathcal{H}$. As a nominal control policy, we choose $a^{\text{dd}}(t) = 0$, which mimics a data-driven controller that does not react to the imminent hazard and therefore does not initiate braking.

The trajectory of the system (15) with the nominal control $a^{\text{dd}}(t)$, supervised by a switching safety filter using $\Omega = \mathcal{S}$, and starting at $(x(0), v(0)) = (-8\ \text{m}, 6\ \text{m s}^{-1})$ is depicted in Figure 5. Once the system reaches the boundary of Ω, the filter intervenes by switching to the backup controller

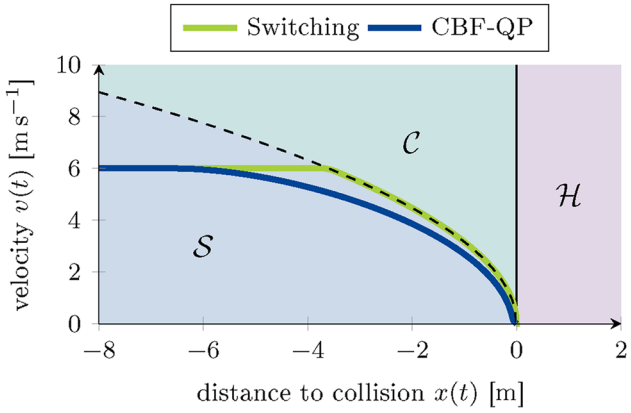

Figure 5: State trajectories in the $(x(t), v(t))$ phase space for a collision-avoidance scenario. The switching safety filter intervenes at the boundary of $\mathcal{S}$, whereas the CBF-QP enforces braking earlier.

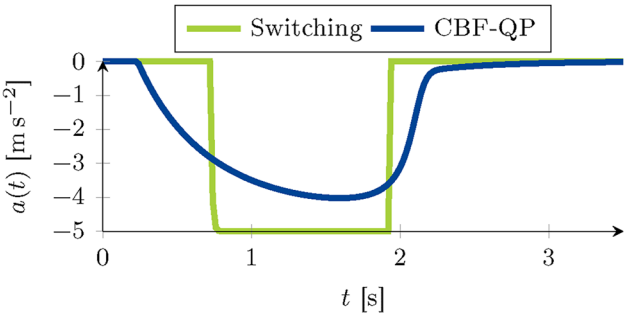

Figure 6: Control inputs for the collision-avoidance example. The switching safety filter results in a hard braking action that saturates at $-a_{\max} = -5$ m s^{-2}, while the CBF-QP yields a smoother deceleration profile by intervening earlier.

$a^{\Omega}(t) = -a_{\max}$. For comparison, Figure 5 also depicts the trajectory of a CBF-QP–based safety filter using $\alpha(h) = 2h$ and

$$h(\bar{x}, \bar{v}) = -\bar{x} - \frac{\bar{v}^2}{2a_{\max}}, \tag{17}$$

which is a feasible CBF for every $(\bar{x}, \bar{v}) \in \mathcal{S}$. The applied control input $\bar{a} \in \mathbb{R}$ is obtained at each time instance by solving a CBF-QP of the form

$$\begin{aligned} \min_{\bar{a} \in \mathbb{R}} \quad & \frac{1}{2}(\bar{a} - a^{\mathrm{dd}}(t))^2 \\ \text{s.t.} \quad & \bar{a} \le -a_{\max} - 2\frac{x(t)}{v(t)} a_{\max} - v(t). \end{aligned} \tag{18}$$

As depicted in Figure 5, the safety filter based on the CBF-QP intervenes earlier than the switching safety filter, i.e., before the boundary of $\mathcal{S}$ is reached. Consequently, the required braking action is less aggressive, which is reflected in the smoother control input depicted in Figure 6. In contrast to the hard switching behavior of the switching safety filter, the CBF-QP avoids abrupt changes in the control input, resulting in reduced jerk while still ensuring safety.

Here, both the switching safety filter and the CBF-based safety filter are based on the maximal safe set $\mathcal{S}$. In practice, however, safety filters often rely on handcrafted invariant sets, which can be overly conservative. Therefore, we now demonstrate how the virtual safe set Ω' of an MPSF expands the admissible operating region of the safety filter in such cases. For this purpose, we use the pragmatic terminal safe set

$$\Omega^{\mathrm{trm}} := \{(\bar{x}, \bar{v}) \in \mathbb{R}^2 \mid \bar{x} \le 0,\ \bar{v} \le 0, \}. \tag{19}$$

We employ a discrete version of the MPSF optimization problem (see (21)) based on the time-discretized dynamics of the double-integrator (15)

$$\begin{aligned} x_{i+1} &= x_i + v_i\, \mathrm{d}t + \frac{1}{2} a_i\, \mathrm{d}t^2, \\ v_{i+1} &= v_i + a_i\, \mathrm{d}t, \end{aligned} \tag{20}$$

with sampling time $\mathrm{d}t \in \mathbb{R}_{>0}$.

The applied control input $a_0 \in \mathbb{R}$ is obtained at each time instance by solving an MPC problem with prediction horizon $N \in \mathbb{N}_{>0}$ of the following form.

$$\begin{aligned} \min_{a_0, \dots, a_{N-1}} \quad & \frac{1}{2}(a_0 - a^{\mathrm{dd}}(t))^2 \\ \text{s.t.} \quad & x_0 = x(t), \quad v_0 = v(t), \\ & \text{for } i = 0, \dots, N-1\text{:} \\ & \quad x_{i+1} = x_i + v_i\, \mathrm{d}t + \frac{1}{2} a_i\, \mathrm{d}t^2, \\ & \quad v_{i+1} = v_i + a_i\, \mathrm{d}t \\ & \quad |a_i| \le a_{\max}, \\ & x_i \le 0, \quad i = 1, \dots, N, \\ & v_N \le 0. \end{aligned} \tag{21}$$

The terminal condition $v_N \le 0$ together with $a_i \ge -a_{\max}$ limits the velocity $v_i \le v_{\max}$ to $v_{\max} = N \cdot a_{\max} \cdot \mathrm{d}t$ such that a feasible braking maneuver exists within the prediction horizon N. Therefore, the maximal braking curve together with the maximum velocity $v_{\max}$ defines the virtual safe set as

$$\Omega' := \{(\bar{x}, \bar{v}) \in \mathcal{S} \mid \bar{v} \le v_{\max}\}, \tag{22}$$

which characterizes the states from which the terminal condition can be satisfied under the imposed constraints.

The trajectory of the system (15) under the nominal control $a^{\mathrm{dd}}(t)$, supervised by the MPSF in (21) with $\mathrm{d}t = 0.05$ s and $N = 30$, and initialized at $(x(0), v(0)) = (-8\text{ m}, 6\text{ m s}^{-1})$, is depicted in Figure 7. The virtual safe set Ω' for

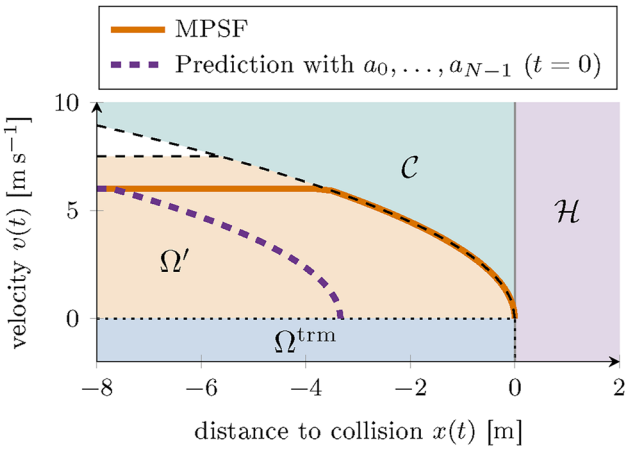

Figure 7: State trajectories in the $(x(t), v(t))$ phase space for a collision-avoidance scenario using an MPSF. At each time step, a feasible trajectory reaching the terminal set Ω^{trm} is computed (the first prediction is shown as a dashed purple line). Braking is initiated at the boundary of the virtual safe set Ω' to ensure convergence to Ω^{trm}, yielding closed-loop behavior similar to that of the least-restrictive switching safety filter, without requiring explicit knowledge of $\mathcal{S}$.

$v_{\max} = 7.5\ \text{m s}^{-1}$ and the predicted trajectory at $t = 0$ are also illustrated. The results demonstrate that the admissible operating region of a conservative terminal set Ω^{trm} can be significantly enlarged by computing, at each instance, a feasible trajectory that reaches the terminal set. As a result, in this example, the closed-loop behavior closely resembles that of the least-restrictive switching safety filter based on $\mathcal{S}$ (cf. Figure 5), but without requiring explicit knowledge of $\mathcal{S}$.

5 Bridging functional safety and data-driven methods through safety filters

Building upon the requirements derived from industrial standards and on the challenges identified for data-driven methods in open environments presented in Section 2, we next examine how safety filters (see Section 4) contribute to these objectives from a functional safety perspective.

5.1 Formal consideration of safety filters within functional safety requirements

To explicitly link functional safety standards to the properties of safety filters, we examine how the operational principles of safety filters relate to the principal requirements and recommendations of the considered standards in Section 2. Table 1 highlights this correspondence that can be addressed directly and unambiguously: risk reduction, and monitoring and intervention at runtime.

Table 1: Relation between the functional safety requirement and the recommendation for data-driven methods formulated in Section 2 and their fulfillment through safety filters.

	Fulfillment by safety filters
Monitoring and intervention (recommendation)	Safety filters continuously observe the system state $\boldsymbol{x}(t)$ and control commands $\boldsymbol{u}^{\text{dd}}(t) = \boldsymbol{\pi}^{\text{dd}}(\boldsymbol{x}(t))$ of the data-driven method to detect deviations from development assumptions. If safety boundaries are violated (i.e., leaving Ω), the intervention strategy switches to a backup control strategy $\boldsymbol{\pi}^{\Omega}$ that guarantees the system remains within the certified safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$. This mechanism implements the required runtime monitoring and intervention within the ODD, ensuring that system behavior remains within the defined safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$.
Risk reduction (requirement)	Safety filters guarantee that a safe control input $\boldsymbol{u}(t) = \boldsymbol{\pi}^{\text{safe}}(\boldsymbol{x}(t), \boldsymbol{u}^{\text{dd}}(t))$ always exists by specifically increasing the controllability (see Definition 3). Consequently, the likelihood that hazardous behavior leads to harm is reduced, lowering the overall risk associated with a hazard. In this way, safety filters directly address the fulfillment of the acceptance criterion A_{H} (see (1)). In an optimal case, the probability that the critical event cannot be controlled $P_{\text{C}\vert\text{E}}$ is formally reduced to 0, provided that the sets $\mathcal{X}$, $\mathcal{U}$, $\mathcal{H}$ are fully known, and accurately measured and a certified safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ exists.

A safety filter continuously monitors the system state $\boldsymbol{x}(t)$ and control commands of a data-driven method $\boldsymbol{u}^{\text{dd}}(t) = \boldsymbol{\pi}^{\text{dd}}(\boldsymbol{x}(t))$ and checks whether they remain within the formally defined safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ (see Definition 11 and Figure 4). Thus, a safety filter detects deviations from the conditions assumed during development and intervenes before states are reached from which it cannot be ensured that harmful states $\mathcal{H}$ will be reached (see $\mathcal{C}$ in Definition 7 and Figure 3). It constitutes a monitoring and intervention mechanism (as recommended in Section 2.5) by identifying potential violations and by overriding or adapting control inputs to keep the system within safe operational bounds represented by Ω (cf. Figures 2 vs. 4).

Additionally, the continuous monitoring and intervention mechanism of the safety filter operationalizes the abstract safety specifications given in Definition 4 in the safe set Ω (see Section 4.3) and links high-level hazard analysis with executable interventions in the control loop, turning safety requirements into formally enforced runtime constraints.

By relying on a monitoring and intervention strategy, safety filters are aligned with the use of formal methods encouraged in IEC 61508 [18] and emphasized in ISO/IEC

TR 5469 [25] (see Section 2.5). Their mathematical formulation provides verifiable safety guarantees and consolidates the safety-critical specifications into the formally defined safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$. This explicit and unified representation of the modeled operating domain and disturbances helps to reduce the practical impact of incomplete specifications (see Challenge 2) by making the underlying safety assumptions understandable. Moreover, the explicit representation of safe sets and the associated proofs enhances the interpretability (see Challenge 6) of the safety argument and makes models transparent and traceable, which is rarely achieved by data-driven controllers. However, the safety guarantee depends on the correct identification of the safety-critical harmful states $\mathcal{H}$ and on the accuracy of the underlying models and measurements. Requirements that cannot be captured within these modeled and measurable bounds remain outside the scope of the safety filter.

The mathematically explicit definition of safe sets $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ makes safety filters suitable for inclusion in safety arguments and structured safety cases required by ISO 26262 [26] and ISO 21448 [27]. This close alignment with formal methods supports treating safety filters themselves as a formal verification and enforcement mechanism. In particular, safety filters offer a means to demonstrate the absence of unreasonable risk by proving the existence and invariance of the certified safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ with respect to the modeled disturbance $\boldsymbol{d}(t) \in \mathcal{D}$ in the defined ODD, thereby reducing the runtime safety argument to a mathematically closed and verifiable problem within these explicitly bounded conditions. Therefore, safety filters support risk reduction by constraining system states to remain within provably safe sets $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$. When demonstrating that the quantitative acceptance criterion A_{H} in (1) is satisfied, the focus lies primarily on the controllability term $P_{\mathrm{C|E}}$, ensuring that hazardous behavior remains within controllable system states. This enables the probability of loss of controllability to be driven arbitrarily close to zero, fulfilling a core objective of functional safety standards by directly helping to keep the overall criterion below the required acceptance criterion A_{H}.

Safety filters further mitigate the effects of data drift (see Challenge 5), not by detecting distributional changes directly, but by monitoring whether the data-driven control policy produces control actions that would drive the system outside the certified safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$.

Finally, we expect safety filters to reduce the testing and certification effort. An important open requirement for their practical use is to substantiate that bounding behavior within a provably safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ can cover whole classes of similar scenarios, including parametrized variations, under a single formal guarantee. Meeting this requirement would allow safety filters to help mitigate excessive test time (see Challenge 3) and compensate for limited data coverage (see Challenge 4) by supporting safety arguments even when not every scenario is explicitly tested. In this way, they could strengthen compliance with functional safety objectives even in unpredictable open environments (see Challenge 1), while naturally remaining constrained to the specified operational domain and the modeling assumptions used for the safety filter design.

5.2 Discussion and practical implications

Having established the fundamental link between safety filters and functional safety requirements, we now consider their practical implications and the limitations that constrain broad industrial adoption.

A particular strength of safety filters lies in complementing data-driven methods and other nominal controllers that lack formal safety guarantees. By translating high-level safety requirements into mathematically verifiable conditions, they can, in principle, reduce the need for exhaustive testing and support the integration of data-driven components into safety-critical applications. However, the actual reduction of verification effort remains an open challenge that must be evaluated in comparison with the established processes prescribed by functional safety standards.

Safety filters address only a defined subset of the overall safety problem. Their guarantees presuppose that the ODD and the relevant hazardous events have first been specified and validated through the HARA framework required by the functional safety standards. Defining and validating the ODD and the set of hazardous events is not a task of the safety filter but must be achieved by other measures within the development process. An open question for practice is whether the explicit and formally verifiable nature of the safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ can provide a quantifiable simplification of these specification and verification activities compared to the procedures currently prescribed by functional safety standards. Once these prerequisites are fulfilled, the safety filter can enforce all-time safety (see Definition 8). Safety filters therefore enforce the formally specified safe sets at runtime but do not replace the processes of ODD definition, hazard identification, or completeness assurance.

At the same time, the review of practical realizations (see Sections 4.3 and 4.4) highlights significant technical limitations. HJ reachability analysis, while least restrictive, is computationally intractable for high-dimensional systems.

CBFs enable efficient online enforcement but lack general constructive mechanisms and may lead to conservative safe sets. MPSFs extend feasibility beyond conservative terminal sets but rely on sufficiently accurate system models and face challenges in meeting real-time constraints. Across all methods, challenges such as chattering near safety boundaries, robustness to uncertainty, and scalability in multi-agent scenarios persist, restricting immediate large-scale industrial adoption, although ongoing research proposes various approaches to mitigate these limitations.

A further challenge lies in the effort required for the formal certification of the safe set. The proof of existence and invariance of the certified control-invariant set Ω, especially the computation of the maximal safe set $\mathcal{S}$ and the verification of its invariance, can, depending on system dimension and model uncertainty, demand a certification effort comparable to that of validating the underlying control function itself.

Furthermore, the assumptions used to derive $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ impose fundamental practical limits. The theoretical link that reduces the controllability term $P_{\mathrm{C|E}}$ to zero presumes that the disturbance set $\mathcal{D}$, the state space $\mathcal{X}$, and the input space $\mathcal{U}$ are completely known and accurately measured. In real systems, however, these sets can only approximate reality: disturbances may be partially unknown or time-varying, sensors introduce stochastic measurement noise, and limited perception can lead to unmodeled effects such as temporary occlusions. Consequently, the computed safe set $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ can guarantee invariance only with respect to the modeled uncertainty, and the probability of entering a state without a safe control option can be assured only within these modeling limitations. Moreover, because $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ itself is often inferred from sensor data and system identification subject to stochastic noise, its boundaries are typically established under probabilistic rather than purely deterministic assumptions. The use of a safety filter as a formal method aims to reduce the probability $P_{\mathrm{C|E}}$ of reaching an uncontrollable state to 0. In practice, this remains an idealization, since limitations in model fidelity, measurement accuracy, and probabilistic set estimation mean that the achievable value of $P_{\mathrm{C|E}}$ can only be reduced.

A quantitative assessment of how explicitly defined control-invariant safe sets $\Omega \subseteq \mathcal{X}_{\mathcal{H}}$ translate into measurable reductions of specification and verification effort compared to established functional safety processes constitutes an important direction for future research.

6 Conclusions

In this paper we have assessed safety filters as a runtime mechanism that ensures safety in data-driven control from the perspective of functional safety as defined in industrial standards. The integration of data-driven control methods into safety-critical systems requires mechanisms that ensure risk reduction in accordance with established functional safety standards. Formally verified safety filters have demonstrated applicability in systems with low-dimensional dynamics and well-structured operating domains, and show strong potential to support testing and certification by constraining the system behavior to provably safe sets. Safety filters can theoretically assure safety by operating within a specified requirement domain and by reducing risk through controllability inside a formally defined invariant safe set. The practical use of safety filters depends on validated assumptions about disturbances and operational conditions and on scalable methods to compute and certify the safe sets. Broad industrial adoption is therefore still limited by the need for clearly defined safety requirements, invariant safety constraints, and progress in computational scalability. Nevertheless, this paper highlights the necessity of provable safety mechanisms that supervise data-driven control and enforce safety through a verified intervention process. Building on requirements derived from industrial functional safety standards and on challenges identified for data-driven methods, this paper recommends the use of formal methods as a means for development and certification processes to ensure functional safety.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: Manuel Hess and Ben-Micha Piscol contributed equally to this work and share first authorship. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: DeepL Translator was used to improve the language of the manuscript.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

[1] V. Tinoco, M. F. Silva, F. N. Santos, R. Morais, S. A. Magalhães, and P. M. Oliveira, "A review of advanced controller methodologies for robotic manipulators," *Int. J. Dyn. Control*, vol. 13, no. 36, 2025, https://doi.org/10.1007/s40435-024-01533-1.

[2] H. Zsifkovits, M. Woschank, S. Ramingwong, and W. Wisittipanich, "State-of-the-art analysis of the usage and potential of automation in logistics," in *Industry 4.0 for SMEs*, D. T. Matt, V. Modrák, and H. Zsifkovits, Eds., Cham, Springer International Publishing, 2020, pp. 193—212.

[3] J. Deichmann, G. Doll, B. Klein, B. Mühlreiter, and J. P. Stein, "Cracking the complexity code in embedded systems development," New York City, McKinsey, 2022.

[4] W. Liu *et al*., "A systematic survey of control techniques and applications in connected and automated vehicles," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21892—21916, 2023.

[5] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Comput. Surv.*, vol. 55, no. 1, pp. 1—36, 2023.

[6] K. Prag, M. Woolway, and T. Celik, "Toward data-driven optimal control: A systematic review of the landscape," *IEEE Access*, vol. 10, pp. 32190—32212, 2022.

[7] U. Rosolia, X. Zhang, and F. Borrelli, "Data-driven predictive control for autonomous systems," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 1, no. 1, pp. 259—286, 2018.

[8] W. Tang and P. Daoutidis, "Data-driven control: Overview and perspectives," in *2022 American Control Conference (ACC)*, 2022, pp. 1048—1064.

[9] S. Mohseni, H. Wang, C. Xiao, Z. Yu, Z. Wang, and J. Yadawa, "Taxonomy of machine learning safety: A survey and primer," *ACM Comput. Surv.*, vol. 55, no. 8, pp. 1—38, 2023.

[10] J. Perez-Cerrolaza *et al*., "Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1—40, 2024.

[11] J. Linnosmaa, P. Tikka, J. Suomalainen, and N. Papakonstantinou, *Machine Learning in Safety Critical Industry Domains (VTT Research Report)*, Tampere, VTT Technical Research Centre of Finland, 2020.

[12] S. Houben *et al*., "Inspect, understand, overcome: A survey of practical methods for AI safety," in *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, T. Fingscheidt, H. Gottschalk, and S. Houben, Eds., Cham, Springer International Publishing, 2022, pp. 3—78.

[13] M. Borg *et al*., "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry," *J. Automot. Software Eng.*, vol. 1, no. 1, pp. 1—19, 2019.

[14] K. Ahmad, M. Abdelrazek, C. Arora, M. Bano, and J. Grundy, "Requirements practices and gaps when engineering human-centered artificial intelligence systems," *Appl. Soft Comput.*, vol. 143, p. 110421, 2023.

[15] A. Corso, R. Moss, M. Koren, R. Lee, and M. Kochenderfer, "A survey of algorithms for black-box safety validation of cyber-physical systems," *J. Artif. Intell. Res.*, vol. 72, pp. 377—428, 2021.

[16] Functional Safety Standards Committee White Paper Working Group, *The Functional Safety Terminology Landscape*, New York, NY, USA, IEEE SA Functional Safety Standards Committee, 2023.

[17] European Commission, Commission Implementing Regulation (EU) 2022/1426 of 5 August 2022 laying down rules for the application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as regards uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles.

[18] International Electrotechnical Commission (IEC), *IEC 61508:2010 Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems*, Geneva, Switzerland, 2010.

[19] R. Engert, P. Zaumseil, D. Zdetski, and T. Brandmeier, "Vehicle safety of autonomous vehicles: Current limitations and potential solutions," *Automatisierungstechnik*, vol. 74, no. 1, pp. 12—34, 2026.

[20] P. Koopman and M. Wagner, "Challenges in autonomous vehicle testing and validation," *SAE Int. J. Transp. Saf.*, vol. 4, no. 1, pp. 15—24, 2016.

[21] K.-C. Hsu, H. Hu, and J. F. Fisac, "The safety filter: A unified view of safety-critical control in autonomous systems," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 7, no. 1, pp. 47—72, 2024.

[22] K. P. Wabersich *et al*., "Data-driven safety filters: Hamilton-Jacobi reachability, control barrier functions, and predictive methods for uncertain systems," *IEEE Control Syst.*, vol. 43, no. 5, pp. 137—177, 2023.

[23] L. Brunke *et al*., "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 5, no. 1, pp. 411—444, 2022.

[24] A. D. Ames *et al*., "Control barrier functions: Theory and applications," in *2019 18th European Control Conference (ECC)*, Naples, Italy, IEEE, 2019, pp. 3420—3431.

[25] International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), *ISO/IEC TR 5469:2024 Artificial Intelligence — Functional Safety and AI Systems*, Geneva, Switzerland, 2024.

[26] International Organization for Standardization (ISO), *ISO 26262:2018 Road Vehicles — Functional Safety*, Geneva, Switzerland, 2018.

[27] International Organization for Standardization (ISO), *ISO 21448:2022 Road Vehicles — Safety of the Intended Functionality (SOTIF)*, Geneva, Switzerland, 2022.

[28] International Electrotechnical Commission (IEC), *IEC 62061:2021 Safety of Machinery — Functional Safety of Safety-Related Control Systems*, Geneva, Switzerland, 2021.

[29] Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery and repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC.

[30] International Organization for Standardization (ISO), *ISO/PAS 8800:2024 Road Vehicles — Safety and Artificial Intelligence*, Geneva, Switzerland, 2024.

[31] International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), *ISO/IEC 22989: Information Technology — Artificial Intelligence — Artificial Intelligence Concepts and Terminology*, Geneva, Switzerland, 2022.

[32] International Organization for Standardization (ISO), *ISO 13849:2023 Safety of Machinery — Safety-Related Parts of Control Systems*, Geneva, Switzerland, 2023.

[33] International Organization for Standardization (ISO), *ISO/SAE PAS 22736:2021 Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, 2021.

[34] X. Zhang *et al*., "Finding critical scenarios for automated driving systems: A systematic mapping study," *IEEE Trans. Softw. Eng.*, vol. 49, no. 3, pp. 991—1026, 2023.
[35] P. Koopman, "The heavy tail safety ceiling," in *Automated and Connected Vehicle Systems Testing Symposium*, 2018.
[36] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?," *Transp. Res. A Policy Pract.*, vol. 94, no. C, pp. 182—193, 2016.
[37] Y. Ma, Z. Wang, H. Yang, and L. Yang, "Artificial intelligence applications in the development of autonomous vehicles: A survey," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 2, pp. 315—329, 2020.
[38] D. Amodei *et al*., *Concrete Problems in AI Safety*, 2016. arXiv: 1606.06565 [cs].
[39] P. Akella *et al*., "Risk-aware robotics: Tail risk measures in planning, control, and verification [focus on education]," *IEEE Control Syst.*, vol. 45, no. 4, pp. 46—78, 2025.
[40] B. Siciliano and O. Khatib, Eds. *Springer Handbook of Robotics (Springer Handbooks)*, Cham, Springer International Publishing, 2016.
[41] A. Khan, W. Xie, B. Zhang, and L.-W. Liu, "A survey of interval observers design methods and implementation for uncertain systems," *J. Franklin Inst.*, vol. 358, no. 6, pp. 3077—3126, 2021.
[42] 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)*, Chicago, IL, USA, IEEE, 2015, pp. 4542—4548.
[43] T. Gurriet, M. Mote, A. D. Ames, and E. Feron, "An online approach to active set invariance," in *2018 IEEE Conference on Decision and Control (CDC)*, Miami, FL, USA, IEEE, 2018, pp. 3592—3599.
[44] M. Althoff and J. M. Dolan, "Online verification of automated road vehicles using reachability analysis," *IEEE Trans. Robot.*, vol. 30, no. 4, pp. 903—918, 2014.
[45] M. Chen *et al*., "FaSTrack: A modular framework for real-time motion planning and guaranteed safe tracking," *IEEE Trans. Automat. Control*, vol. 66, no. 12, pp. 5861—5876, 2021.
[46] S. Kousik, S. Vaskov, F. Bu, M. Johnson-Roberson, and R. Vasudevan, "Bridging the gap between safety and real-time performance in receding-horizon trajectory design for mobile robots," *Int. J. Robot Res.*, vol. 39, no. 12, pp. 1419—1469, 2020.
[47] J.-P. Aubin, A. M. Bayen, and P. Saint-Pierre, *Viability Theory: New Directions*, Berlin, Heidelberg, Springer Berlin Heidelberg, 2011.
[48] M. Nagumo, "Über die Lage der Integralkurven gewöhnlicher Differentialgleichungen," *Proc. Phys.-Math. Soc. Jpn. 3rd Ser.*, vol. 24, pp. 551—559, 1942.
[49] R. Isaacs, *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*, Mineola, NY, USA, Dover Publications Inc., 1999.
[50] 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)*, Melbourne, VIC, Australia, IEEE, 2017, pp. 2242—2253.
[51] C. Bohn, M. Hess, and S. Hohmann, *Captivity-Escape Games as a Means for Safety in Online Motion Generation*, 2025. arXiv: 2506.01399v2 [eess].
[52] I. Mitchell, A. Bayen, and C. Tomlin, "A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games," *IEEE Trans. Automat. Control*, vol. 50, no. 7, pp. 947—957, 2005.
[53] M. Chen, S. L. Herbert, M. S. Vashishtha, S. Bansal, and C. J. Tomlin, "Decomposition of reachable sets and tubes for a class of nonlinear systems," *IEEE Trans. Automat. Control*, vol. 63, no. 11, pp. 3675—3688, 2018.
[54] S. L. Herbert, S. Bansal, S. Ghosh, and C. J. Tomlin, "Reachability-based safety guarantees using efficient initializations," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, Nice, France, IEEE, 2019, pp. 4810—4816.
[55] C. Bohn, P. Reis, M. Schwartz, and S. Hohmann, "Time and memory-efficient computation of Hamilton-Jacobi reachable sets based on a level set method employing adaptive grids," presented at the 2023 62nd IEEE Conference on Decision and Control (CDC), Singapore, Singapore, IEEE, 2023, pp. 8235—8241.
[56] J. H. Gillula, H. Huang, M. P. Vitus, and C. J. Tomlin, "Design of guaranteed safe maneuvers using reachable sets: Autonomous quadrotor aerobatics in theory and practice," in *2010 IEEE International Conference on Robotics and Automation*, Anchorage, AK, USA, IEEE, 2010, pp. 1649—1654.
[57] A. Bajcsy, S. Bansal, E. Bronstein, V. Tolani, and C. J. Tomlin, "An efficient reachability-based framework for provably safe autonomous navigation in unknown environments," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 1758—1765.
[58] C. Bohn, J. Bosch, M. Hess, and S. Hohmann, "Reducing conservatism in fast and safe motion generation by means of captivity-escape games," in *2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC)*, 2025, pp. 1518—1524.
[59] T.-W. Hsu, J. J. Choi, D. Amin, C. Tomlin, S. C. McWherter, and M. Piedmonte, "Towards flight envelope protection for the NASA tiltwing eVTOL flight mode transition using Hamilton-Jacobi reachability," *J. Am. Helicopter Soc.*, vol. 69, no. 2, pp. 1—18, 2024.
[60] B. Li, S. Wen, Z. Yan, G. Wen, and T. Huang, "A survey on the control lyapunov function and control barrier function for nonlinear-affine control systems," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 3, pp. 584—602, 2023.
[61] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Trans. Automat. Control*, vol. 62, no. 8, pp. 3861—3876, 2017.
[62] K. Garg *et al*., "Advances in the theory of control barrier functions: Addressing practical challenges in safe control synthesis for autonomous and robotic systems," *Annu. Rev. Control*, vol. 57, p. 100945, 2024.
[63] Z. Jarvis-Wloszek, R. Feeley, W. Tan, K. Sun, and A. Packard, "Control applications of sum of squares programming," in *Positive Polynomials in Control*, D. Henrion and A. Garulli, Eds., Berlin, Heidelberg, Springer Berlin Heidelberg, 2005, pp. 3—22.
[64] A. Majumdar and R. Tedrake, "Funnel libraries for real-time robust feedback motion planning," *Int. J. Robot Res.*, vol. 36, no. 8, pp. 947—982, 2017.
[65] X. Xu, J. W. Grizzle, P. Tabuada, and A. D. Ames, "Correctness guarantees for the composition of Lane keeping and adaptive cruise control," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 3, pp. 1216—1229, 2018.
[66] P. A. Parrilo, "Semidefinite programming relaxations for semialgebraic problems," *Math. Program.*, vol. 96, no. 2, pp. 293—320, 2003.
[67] S. Tonkens and S. Herbert, "Refining control barrier functions through Hamilton-Jacobi reachability," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, IEEE, 2022, pp. 13355—13362.

[68] J. J. Choi, D. Lee, K. Sreenath, C. J. Tomlin, and S. L. Herbert, "Robust control barrier—value functions for safety-critical control," in *2021 60th IEEE Conference on Decision and Control (CDC)*, Austin, TX, USA, IEEE Press, 2021, pp. 6814—6821.
[69] X. Xu, P. Tabuada, J. W. Grizzle, and A. D. Ames, "Robustness of control barrier functions for safety critical control," *IFAC-PapersOnLine*, vol. 48, no. 27, pp. 54—61, 2015.
[70] P. Jagtap, G. J. Pappas, and M. Zamani, "Control barrier functions for unknown nonlinear systems using gaussian processes," in *2020 59th IEEE Conference on Decision and Control (CDC)*, Jeju, Korea (South), 2020, pp. 3699—3704.
[71] J. J. Choi, D. Lee, K. Sreenath, C. J. Tomlin, and S. L. Herbert, "Robust control barrier—value functions for safety-critical control," in *2021 60th IEEE Conference on Decision and Control (CDC)*, Austin, TX, USA, IEEE, 2021, pp. 6814—6821.
[72] M. Jankovic, "Robust control barrier functions for constrained stabilization of nonlinear systems," *Automatica*, vol. 96, pp. 359—367, 2018.
[73] M. H. Cohen, C. Belta, and R. Tron, *Robust Control Barrier Functions for Nonlinear Control Systems with Uncertainty: A Duality-based Approach*, 2022. arXiv: 2208.05955 [math].
[74] S. Kolathaya and A. D. Ames, "Input-to-state safety with control barrier functions," *IEEE Control Syst. Lett.*, vol. 3, no. 1, pp. 108—113, 2019.
[75] A. Alan, A. J. Taylor, C. R. He, G. Orosz, and A. D. Ames, "Safe controller synthesis with tunable input-to-state safe control barrier functions," *IEEE Control Syst. Lett.*, vol. 6, pp. 908—913, 2022.
[76] S. Prajna, A. Jadbabaie, and G. J. Pappas, "A framework for worst-case and stochastic safety verification using barrier certificates," *IEEE Trans. Automat. Control*, vol. 52, no. 8, pp. 1415—1428, 2007.
[77] A. Clark, "Control barrier functions for stochastic systems," *Automatica*, vol. 130, p. 109688, 2021.
[78] Y. Lyu, W. Luo, and J. M. Dolan, "Probabilistic safety-assured adaptive merging control for autonomous vehicles," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China, IEEE, 2021, pp. 10764—10770.
[79] R. Namerikawa, A. Wiltz, F. Mehdifar, T. Namerikawa, and D. V. Dimarogonas, "On the equivalence between prescribed performance control and control barrier functions," in *2024 American Control Conference (ACC)*, 2024, pp. 2458—2463.
[80] L. Lanza, J. Köhler, D. Dennstädt, T. Berger, and K. Worthmann, "A model-free approach to control barrier functions using funnel control," *IEEE Control Syst. Lett.*, vol. 9, pp. 1183—1188, 2025.
[81] K. P. Wabersich and M. N. Zeilinger, "Linear model predictive safety certification for learning-based control," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 7130—7135.
[82] 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.
[83] K. Wabersich, L. Hewing, A. Carron, and M. Zeilinger, "Probabilistic model predictive safety certification for learning-based control," *IEEE Trans. Automat. Control*, no. 1, p. 1, 2021.
[84] W. S. Cortez, J. Drgona, D. Vrabie, and M. Halappanavar, *A Robust, Efficient Predictive Safety Filter*, 2024. arXiv: 2311.08496 [eess].
[85] A. Didier, A. Zanelli, K. P. Wabersich, and M. N. Zeilinger, "Predictive stability filters for nonlinear dynamical systems affected by disturbances," *IFAC-PapersOnLine*, vol. 58, no. 18, pp. 200—207, 2024.
[86] E. Milios, K. P. Wabersich, F. Berkel, and L. Schwenkel, "Stability mechanisms for predictive safety filters," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, 2024, pp. 2409—2416.

Bionotes

Manuel Hess
Institute of Control Systems (IRS), Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
manuel.hess@kit.edu

Manuel Hess received the B.Sc. degree in Mechatronics from Hamburg University of Technology (TUHH), Hamburg, Germany, and the M.Sc. degree in Electrical Engineering and Information Technology from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, including a semester abroad at the University of Porto (FEUP), Porto, Portugal. Since 2023, he has been a Research Scientist in the research group Functional Safety Control at the Institute of Control Systems (IRS), KIT. His research interests include control of safety-critical systems with formal safety guarantees.

Ben-Micha Piscol
Institute of Control Systems (IRS), Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
ben-micha.piscol@kit.edu

Ben-Micha Piscol received his B.Eng. degree in Electrical Engineering from the Baden-Wuerttemberg Cooperative State University (DHBW) Stuttgart in cooperation with Robert Bosch GmbH, and the M.Sc. degree in Electrical Engineering and Information Technology from the Karlsruhe Institute of Technology (KIT), Germany. Since 2022, he has been a Research Scientist in the research group Functional Safety Control at the Institute of Control Systems (IRS), KIT. His research focuses on the safety assurance of data-driven control in autonomous systems, with an emphasis on the systematic integration of safety filter concepts into safety cases.

Christopher Bohn
Institute of Control Systems (IRS), Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
christopher.bohn@kit.edu

Christopher Bohn received his Master's degree in Electrical Engineering and Information Technology from the Karlsruhe Institute of Technology (KIT), Germany, in 2020. Since 2020, he has been a Ph.D. student with the Institute of Control Systems at KIT under the supervision of Prof. Sören Hohmann. His research interests include adaptive robust motion generation, with a focus on ensuring safe system behavior in the presence of model mismatch, disturbances, and system failures.

Sören Hohmann
Institute of Control Systems (IRS), Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
soeren.hohmann@kit.edu

Sören Hohmann received the Diploma and Ph.D. degrees in Electrical Engineering from the University of Karlsruhe (now Karlsruhe Institute of Technology (KIT)), Karlsruhe, Germany, in 1997 and 2002, respectively, after studying electrical engineering jointly at the Technische Universität Braunschweig, the University of Karlsruhe, and the École Nationale Supérieure d'Électricité et de Mécanique, Nancy, France. Afterwards, until 2010, he worked in the industry for BMW Munich, Germany, in various positions, where his last position was the head of the predevelopment and series development of active safety systems. He is currently the Head of the Institute of Control Systems (IRS), KIT, Karlsruhe, and the Director's Board Member of the Research Center for Information Technology (FZI), Karlsruhe. His research interests are cooperative control, control of networked and multi-energy systems, and system guarantees by design.