To We Robot 2023 Participants:

Thank you for reading! This is still a draft. We appreciate any questions, suggestions, feedback, or comments.
- Kegan Strawn

Certifiable Safety Techniques in Mobile Robots as Tools for Precise and Assistive AI Regulation

Kegan J. Strawn[1] and Daniel Sokol[2]

[1]Department of Computer Science, University of Southern California, Los Angeles, California, USA.
[2]School of Law, University of Southern California, Los Angeles, California, USA.

Corresponding author: kegan.j.strawn@usc.edu;

Abstract

The widespread application of artificial intelligence (AI) systems across industries such as healthcare, finance, and technology has surged, raising complex legal and regulatory challenges. Adapting to and addressing bias, discrimination, and accuracy issues as technology progresses rapidly proves difficult. Deploying autonomous vehicles, delivery robots, and other physically embodied AI agents in pedestrian environments adds new layers of complexity for regulation. This paper examines recent trends in regulating AI and autonomous vehicles and current research in explainability, validation, and statistical verification of learning-based autonomous systems to emphasize the importance of diving deep into engineering techniques to produce well-informed and nuanced regulatory approaches and encourage interdisciplinary work. Using a collision avoidance example, safety and validation methods are explored to bolster confidence in robot deployment and suggest a more context-specific approach to regulation. These tools can ensure public acceptance, performance accountability, third-party assessment, and adaptive regulation. We highlight in the example the ongoing challenge of safe interactions with dynamic agents and the need for regulatory guidance in balancing innovation and public safety. We add to the growing calls for interdisciplinary discussions among roboticists, computer scientists, legal professionals, policymakers, and ethics experts to generate informed regulations that promote the development of AI technologies that operate safely under uncertainty while supporting innovation.

Keywords: Robotics, Artificial Intelligence, Safety, Regulation, Autonomous Vehicles

1 Introduction

Implementing artificial intelligence (AI) systems in industries such as healthcare, financial, social, and other online systems has grown exponentially in recent years. AI introduces important legal and regulatory questions when confronting relatively black-box AI models. Integrating AI into various aspects of daily life, particularly in the form of relatively opaque AI models, has brought forth many complex legal and regulatory challenges that are actively being discussed, proposed, and critiqued [1, 2]. Potential harm from bias, discrimination, and inaccuracy is difficult for regulatory bodies to address. These issues are compounded when applied to autonomous vehicles, delivery robots, and other AI-driven agents operating within pedestrian environments [3, 4]. Robotic systems, such as autonomous vehicles, are already being deployed before the non-embodied AI challenges have been resolved, introducing additional challenges and pressure for regulation to keep pace with technological advancement. One pressing concern revolves around the inherent opaqueness of certain AI models, which makes it difficult to understand and explain their decision-making processes [5]. This lack of transparency hampers accountability, a critical aspect of regulating these technologies.

This paper discusses the role of explainability, validation, and statistical guarantees in learning-based autonomous systems, offering them as tools for certifying embodied AI and accountability. With AI's rapid proliferation, addressing issues of bias and discrimination has become imperative, compounded by the complexities introduced by physically embodied AI in pedestrian environments. This paper presents safety and validation methods and highlights innovative approaches, such as probabilistic safety filters, to enhance confidence in robot deployment and call for more nuanced regulation solutions that consider the modularity and complexity of each AI application. Similar methods could support regulatory tools, including certification and probabilistic guarantees, to contribute to public assessment, performance accountability, and informed regulation, fostering a balanced approach to robotic innovation and human safety.

Despite impressive progress in autonomous vehicles, robot, and multi-robot navigation research throughout the last two decades, negotiating safe interactions with other dynamic agents remains a fundamental challenge [6]. Understanding other agents' intents and policies is critical for safety systems. However, these intents are often unknown. Communication can resolve some collision conflicts, but communication may be unavailable or not permitted for many robotics applications. Therefore, a decentralized collision avoidance planning algorithm is necessary as it can reduce collisions in these scenarios. Still, it may be flawed and risks negative interactions or harming the other agents. Navigating around pedestrians and other robots or vehicles safely requires regulatory pressure to ensure the production of robots that maximize aid and minimize harm. In the ever-evolving landscape of autonomous vehicles, robotics, and multi-robot navigation, the persistent challenge of ensuring safe interactions with dynamic agents remains paramount.

Moreover, deploying AI systems in these contexts raises serious questions about safety and ethics. The ability of AI-driven agents to navigate around pedestrians, other robots, and vehicles is still an ongoing engineering challenge. This challenge is not only technical but also regulatory in nature. Regulators must strike a balance between promoting innovation and safeguarding public safety. They must establish guidelines and standards that ensure these AI-powered systems prioritize the well-being of pedestrians and minimize harm. Additionally, determining liability in accidents involving AI-driven agents is a complex legal matter that requires careful consideration [7].

This paper first provides background on explainability, validation, and statistical guarantees in AI and AI for autonomous vehicles. We explain existing safety research, including control-barrier functions, reachability analysis, and statistical guarantees. This includes key differences in machine learning models for general AI and machine learning for physically embodied AI. Then, we highlight an example in current research in robotic safety and validation as a suggestion for nuanced regulation and policy for complex embodied AI. Robots in pedestrian environments pose a significant physical risk to humans and other agents. We present robotic safety and validation methods to establish public safety, trustworthy products, and safe policies. To understand the current safety strategy, we outline popular safety techniques for neural network-based motion controllers in the presence of dynamic pedestrians with uncertain trajectories and behaviors. Setting AI into specific scenarios helps generate well-informed regulation that can adapt to the needs of the diverse problems AI can be applied to and the unique methods behind each AI application.

We leverage recent proposals for AI regulation in combination with research in robotic safety to ignite inclusive and nuanced discussions around approaches that balance robotic regulation and safety without restricting innovation. The regulation of AI in human-in-the-loop physical systems, such as autonomous vehicles and delivery robots, demands a multifaceted approach. It necessitates addressing issues of transparency, accountability, safety, and ethics to foster the development of technologies that can assist and coexist harmoniously with humans in pedestrian environments. We invite further dialogue from interdisciplinary researchers across robotics, computer science, law, and ethics that builds from these tools. We aim to spark informed regulation for bettering these robotic systems, those who develop them, and those they affect.

2 Regulation of AI

Implementation of AI systems in industries such as healthcare, financial, social, generative, and other online systems has grown in recent years. Potential harm from bias, discrimination, and inaccuracy is difficult for regulatory bodies to address but has received increasing attention. AI introduces important legal and regulatory questions when confronting relatively black-box AI models.

First, it is important to define safety and who safety is being defined for and by, as safety should be defined in ways that motivate equitable uses of automated technologies and limit uses that perpetuate harm or violence [8]. Throughout history and in recent work, "public safety" in government regulation has failed to make all people safe, and without careful work, some public members can be unsafe [8]. It will be as important to define what goals and values we want the system to align with as it is to develop the safety critical systems themselves. Leaving this choice up to only the developers could directly or indirectly fail to build safe systems for everyone. The Oxford English Dictionary defines "safety" as "the state of being safe" and as "the state of being protected from or guarded against hurt or injury; freedom from danger." The National Institute of Standards and Technology (NIST) defines safety for an AI system as a system that should "not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered." This work will focus on safety regarding a collision avoidance example for autonomous mobile robots and vehicles. Here, safety is naturally defined as limiting the chance of physical collision. By focusing on the specific application of AI, we can advance this discussion further than other previous work. However, more broadly, safety for autonomous robots and vehicles will require future work to define safety in the most equitable and informed definition to help align the intentions of our safety techniques with our impact on society and our communities.

When considering risk assessment and regulation of AI technologies, as we will in this paper, it is important to consider whether existing structures and procedures can suit and adapt to the problem or if the problem is entirely new and requires a standalone solution. In this paper, we suggest tools that could be useful to either approach and generally fit the definition of risk assessment, mitigation, and prevention. Any solutions or structures must be adaptable as technology continues to change. We refer to recent work proposing that regulation can take different forms, exemplified by four models of existing risk regulation, each with divergent goals and methods [2]. Different parties of a specific industry area or application may call for different models of risk regulation that lead to conflicting outcomes. The approaches could include quantitative analysis, democratic oversight, regulatory capacity and enforcement, or enterprise risk management [2]. Risk assessment and mitigation typically require developers to mitigate risks and conduct risk analysis rather than banning technology or conditioning its use on specific licensing requirements. It can involve analyzing a system to identify what risks it raises, mitigating any identified risks, testing to ensure risk mitigation is effective, and iteratively going back through the steps as time passes [9]. The more common licensing is a precautionary regulation where the expert agency has set standards for and assesses the safety and performance of a technology before it is used. Another precautionary regulation is sandboxing, where the technology is permitted subject to regulatory supervision within an enforcement safe harbor [10], allowing the development of what might otherwise be banned. We will provide examples of safety validation that could be used in any of these risk assessment styles.

A combination of self-regulation and supervisory enforcement is increasingly necessary for robotics. Recent work calls for pre-deployment risk assessments, external scrutiny of model behavior, risk assessments to inform deployment decisions, and monitoring of new information about model capabilities post-deployment performance [11]. Contributing to a growing need for explainable and certifiable AI techniques to facilitate these goals. Engineering projects should consider these issues from the beginning of the project. We note that risk regulation works best on quantifiable problems and is more challenging for qualitative analysis. When possible, we must examine and explain AI from quantitative means, but for a deep and comprehensive approach, we should explore the performance of AI and the goals built into AI beyond quantitative measurements or find ways to turn them into quantitative analysis when possible.

2.1 Government Guidance

U.S. President Joe Biden recently met with AI leaders to discuss AI's potential and risk [1]. At the forum, AI was framed as America's new moonshot mission, with suggestions for the government to invest in AI as it has with NASA and to adopt a mindset toward an American-led AI achievement urgently. Part of the plan would include increasing access to AI research and resources for academics and industry leaders, such as the National AI Research Resource (NAIRR). The forum covered multiple risks of regulating AI, centered around stifling innovation with uninformed regulation. The fear of overregulation is extremely present in tech communities; the government is concerned about losing its competitive advantage on AI to another country, and many researchers worry about reducing the potential for extraordinary and potentially life-saving breakthroughs. While others, such as the recent FTC report, warned the U.S. Congress about AI, urging policymakers to exercise "great caution." The report outlines concerns on inaccurate, biased, and discriminatory AI tools [12].

As we consider the ways to regulate embodied AI, it is important to consider the regulatory landscape around general AI first. In October of 2022, the White House Office of Science and Technology Policy released the ”Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People," identifying five principles that should guide the design, use, and deployment of AI systems to protect the American public and advance the Biden administration's principles for civil rights and democratic values at risk with algorithmic decision-making and privacy concerns. The blueprint's framework suggests high-level protections to be applied to all automated systems that can meaningfully impact individuals' or communities' exercise of their rights, opportunities, or access. The summarized suggestions call for developing diverse teams, proactively addressing bias and discrimination, ensuring privacy as a default, providing explanations of the system's reasoning, and the ability to fall back on human operators or to appeal a decision. The original goal for the blueprint was for a bill of rights, but the project ended as a whitepaper and a framework of high-level suggestions, which critics argue is much less ambitious, non-binding, non-enforceable, vague, and flawed [13]. However, the blueprint highlights considering the government's role and progress in regulating AI and presents a useful framework to consider the important principles of regulating AI.

The first principle for protection from unsafe or ineffective systems calls for diverse communities, stakeholders, and experts to be consulted and integrated into the team when developing the systems. A system should undergo pre-deployment testing, risk identification, and mitigation with ongoing monitoring. The systems should be designed with safe goals that do not endanger people and can be evaluated independently with the ability to confirm the system's safety. The second principle asserts that people should not face discrimination by algorithms, and systems should be used and designed equitably. Algorithms that can affect or use a protected class (and its proxies) for reasoning should be proactively designed to counteract the potential for bias and discrimination. The third principle calls for data privacy and that users should be protected from privacy violations through design choices that ensure protections by default and provide the user agency over how their data is used or sold. The fourth principle suggests that automated systems must be understandable to the users they impact. Automated systems should provide technically valid, meaningful, and useful explanations–calibrated to the level of risk based on the context. Arguing for reporting that includes summaries of information in plain language and clarity of assessments provided to the public whenever possible. The final principle states that human alternatives and the ability to opt out should be possible where appropriate, with the possibility of appealing an automated decision, especially in high-risk settings such as criminal justice systems. As we generally consider autonomous vehicles and mobile robots, we should consult the five principles to help spot issues and challenges.

Another American bill, Washington Bill SB 5116, would govern the government procurement, development, and use of automated decision systems [14], proposing many of the same risk mitigation techniques and adding an accountability report. The Algorithmic Accountability Act of 2022 would require certain companies to conduct algorithmic impact assessments and risk mitigation for dangerous or critical decisions [15].

The Department of Defense's (DoD) Test & Evaluation, Validation, and Verification (TEVV) aims to help guide the development and adoption of unmanned or semi-autonomous systems [16]. Their focus is to develop approaches for certification that are sufficiently predictive of performance to build critical trust in the systems necessary to deploy at scale. The DoD emphasizes that perfection is not the goal, and only allowing near-perfect results will hurt innovation and the ability to field the systems. Understanding and trustworthiness by the operator are more important, with accurate predictions of performance being given and assessed.

Outside of the U.S., the European Union General Data Protection Regulation (GDPR) aims to provide more control over individuals' personal data, setting guidelines on the explanation and decisions made based on users' data. The GDPR requires personal data processors to provide the rationale behind, or the criteria relied on in reaching a fully automated decision. The current regulations focus mainly on user data protection and privacy, but it is expected to expand into algorithmic transparency and explanation requirements for AI systems [17]. The impact of the GDPR, now multiple years old, is still uncertain but likely to be major [18].

2.2 Explainable Systems

Work in explainable AI systems is highlighted as a possible solution towards accountable AI, making it possible for end users and regulators to reason about the decision-making process [5]. Algorithmic transparency and explainability can provide user awareness [5]; bias and discrimination detection [19, 20]; interpretable behavior of intelligent systems [21]; and accountability for users [22]. When considering the growing body of examples of discrimination and other legal aspects of algorithmic decision-making, researchers call for AI's transparency and accountability under the law to mitigate adverse effects [5, 23–25]. However, these explainable techniques are still in the early stages of research and will require more work to find the right balance between open-sourcing the reasoning behind a system and allowing companies to keep key trade secrets while not limiting the system's performance or providing inaccurate explanations. Although recent work claims it is technically feasible to extract the kinds of explanations currently required of humans from AI systems [23], it is important to note that not all explainability techniques can be applied broadly to all forms of machine learning or AI. The work explores how we can take advantage of what AI systems can bring while also holding them accountable. However, each AI algorithm might require a unique form of explainability.

Accountability is associated with being called 'to account' to an authority for one's actions and is related to responsibility and liability [26, 27]. Accountability in the human and machine context determines whether a system's decision was made in compliance with procedural and substantive standards and holds a party responsible when failing to meet the standards. This can be difficult for autonomous vehicles, as the vehicle is composed of multiple sensors and systems manufactured by different companies and with various techniques that likely include multiple methods of optimization and learning [23, 28]. For accountability, the main focus is on explanations. Asking questions to find the main factors in a decision, what would a change in a certain factor have changed the decision, and why did two similar-looking cases get different decisions [23]. The interested parties for explanation may be end-user decision-makers, affected users, regulatory bodies, or AI system builders. Accountability is important for stakeholders who seek confidence in the system, and explainability can help increase trust and understanding in AI-based systems.

We note that some areas of previous work in regulating AI seem separate from the regulation of mobile robots, such as fair use of generated content. However, even in copyright concerns, scholars suggest clearer legal standards and a robust technical agenda that includes filters on the input and output data to recognize when AI is pushing the boundaries too far or constraining the training of models with legal principles such as fair use in mind. Doing so inevitably requires the collaboration of legal and technical experts to generate AI and policy that aligns with societal values [29]. Effective collaboration between policymakers and technical AI researchers should increase, explore, forestall, and alleviate the risks associated with AI's potential applications [30]. Lessons can be taken from other fields of computer science, such as computer security and data privacy, that have historically gained more attention. Researchers recommend prioritizing inclusivity by involving a broad spectrum of stakeholders and domain experts when discussing or working on these challenges [30]. To help support this and to provide actionable tasks, it is recommended that auditing algorithms and models play a role in investigating any presence of bias or other issues in ways without requiring specific design details. Additionally, ad-hoc explainers and statistical guarantees that may not describe an entire model or method but can be added to improve certain system safety certifications and explanations are interesting for regulatory purposes [31–34]. Approaches that focus on the data input or the results and output of the application, or explainable and accountable algorithms that help clarify the safety and why a certain decision was made without revealing too much information about the system itself will be useful and necessary to foster beneficial AI applications [5, 24, 35].

We reiterate the need for cross-discipline collaboration, as researchers in different disciplines focus on diverse objectives and can have independent versions of explainable AI research, posing a challenge for identifying appropriate design and evaluation methodology across efforts [5]. The previous work shows a trend in labeling AI broadly and proposing non-specific recommendations for fair and safe AI. Many discussions in the previous work do not use examples or show use cases. However, AI is an umbrella term for countless algorithmic approaches with unique capabilities, strengths, risks, and limitations. Accurately discussing AI requires placing AI in the context of its application and the specific AI technique.

3 Mobile Robots

Robotics research works towards developing and deploying systems that can operate autonomously in complex scenarios, often close to humans, and with uncertainty in the sensed environment or beliefs of other agents' intentions in the environment. Examples of robotics applications include self-driving vehicles [36], aerial delivery [37], and service tasks [38]. This paper will generally focus on mobile robots that could serve multiple tasks and, more specifically, how it applies to autonomous vehicle regulation.

Americans drive nearly 3 trillion miles in a single year [39]. This time spent driving increases emissions and is widely considered a dangerous risk [4]. In 2022, there were $42,795$ fatal crashes in the U.S. [40]. It has been reported that 90% of all car accidents are estimated to be caused by human errors [40, 41]. Removing human error from driving and increasing the adoption of fleets of autonomous vehicles, with probabilistic models optimizing the traffic flow and the improved local performance of individual vehicles, could save tens of thousands of lives [4, 42, 43]. Not only could lives be saved if autonomous vehicles are widely adopted, additional benefits include reduced fuel consumption [44, 45], less pollution and improved traffic flow [44, 46, 47], greater mobility for all ages and physical capabilities [44], and increased productivity [4, 44]. However, this assumes the safe and satisfactory performance of autonomous vehicles. Autonomous vehicles and mobile robots are not without risks and will require critical work in regulation and development for safe autonomous robots [48]. Guaranteeing the safety of a complex system like autonomous vehicles and other robots is challenging as the technology behind them is increasing in complexity to take on more complex tasks, are commonly composed of many inter-connected sub-systems, and make use of nonlinear neural networks within these systems [4]. Without knowing how a system will perform, it may never be able to deploy in the real world, or if it is deployed and the actions cannot be certified and explained, then the benefits may be outweighed by association with any negative outcomes. It is in all parties' interests to demonstrate, explain, and certify the safety of actions taken by mobile robots, such as autonomous vehicles.

4 AI and Safety in Mobile Robots

Radio Corporation of America Laboratories is considered to have successfully tested the first radio-operated vehicle in 1921. In 1953, it was followed by a miniature vehicle navigating via wires. Limited by the technology then, the breakthroughs in autonomous robotics began in the 1980s and 1990s thanks to the advancement of computing power. Early efforts were largely confined to controlled environments, such as laboratories and closed test tracks. However, it was not until the 21st century that we witnessed a significant paradigm shift. The Autonomous Land Vehicle program by the US Defence Advanced Research Projects Agency (DARPA), in collaboration with Carnegie Mellon University, Stanford University, and other academic institutions, helped generate methodologies such as the integration of LiDAR, computer vision, and automated control methods still used in autonomous vehicles today. Advances in sensor technology, machine learning, and computing power converged to enable vehicles to navigate real-world, unstructured environments. The DARPA Grand Challenges, where autonomous vehicles navigated challenging desert terrains, and the proliferation of tech companies like Google, Uber, and Tesla that invested heavily in autonomous driving technology helped progress to today, where we stand at the precipice of a transformative era in transportation. The gains in both software and hardware are reshaping how we envision the future of mobility [49].

Today, mobile robots such as autonomous vehicles encompass many research areas and interdisciplinary work, with AI used in an end-to-end fashion and within subcategories such as localization, static object detection, dynamic object detection, scene understanding, tracking, prediction, planning, control, hardware and software considerations, simulation technologies, interpretability, human-machine interactions, and testing. Some of these tasks are dependent on the others. For example, the planning system will be dependent on the sensor data received, the world map that has been constructed, any communications between the robot and others, and the belief of the other agent's intentions. Noise or error could be present in all of these, leading to increasing complexity and uncertainty in the decision-making process [50]. For example, one of the largest tasks for an autonomous mobile robot or vehicle is to plan safe trajectories to move throughout the environment. Most traditional methods to compute safe trajectories for autonomous vehicles are based on one of three approaches [51, 52]: (1) input space discretization with collision checking that can be effective but overly simple or difficult to compute [53–55], (2) randomized planning for the probabilistic exploration of state spaces at a high computational cost [56, 57], (3) constrained optimization and receding-horizon control for computing collision-free trajectories to avoid other traffic with smooth trajectories and direct encoding of the vehicle model but can be complex and settle in local optima [58–60]. We refer to a survey paper for a detailed overview of the perception, localization, segmentation, planning, and end-to-end methods [4].

There are three types of collaborative autonomy: (1) series autonomy, where the human orders a task for the robot to execute; most autonomous approaches fall under this category; (2) interleave autonomy, where the user and the robot take turns operating, and (3)parallel autonomy, or shared control, where the autonomous system functions as a filter or supervisor in the background to ensure safety of the user operator. Different elements of autonomous vehicles and mobile robots can fall under these three categories. However, we will focus on autonomous mobile robots with the intent of series autonomy, where the robot executes tasks on behalf of the robot.

Machine learning methods show great promise in providing excellent performance for complex and non-linear control problems and generalizing previously learned rules to new scenarios [41]. Although we note that while good at generalizing compared to classical methods, it is important to consider the data and training environment for any learning-based solution and recognize that the deployed system will likely encounter examples it has not seen before. These shifts in data distribution seen offline during training and online during execution require new techniques to ensure the safety of learning-based approaches. Most learning for autonomous vehicles and robots can be categorized into supervised and reinforcement learning. The objective is to update the weights of a network such that it can represent a useful function for its task. Supervised learning requires labeled data or demonstrations, and reinforcement learning requires trial and error. For safety in supervised learning, the dataset and the produced policy must be analyzed [61, 62]. In reinforcement learning, the simulator or learning environment and the produced policy are analyzed for safety [3]. In control theory, a model-driven approach leverages the dynamics model to guarantee known environment and agent conditions and information to formalize the problem correctly. Recent trends combine these approaches to gain generalizability and safety within defined boundaries.

The number of contributions to safe learning methods for real-world robotic deployments from control and reinforcement learning research communities has increased significantly over the last ten years [34].

Approaches can apply soft (level one safety), probabilistic (level two safety), or hard (level three safety) constraints. Constructing these constraints can be difficult. We will later look more closely at certifying the safety of learned mobile robot systems at levels two and three in the context of an example problem, solution, and application.

Adaptive control considers the system with uncertain parameters and adapts the controller or model online to optimize performance, requiring knowledge of the parametric form of the uncertainty and relying on a specific model structure. With limitations such as overfitting to the latest observations and no guarantees of convergence, recent work uses learning-based adaptive control approaches [63]. Robust control guarantees stability for prespecified bounded disturbances, finding a suitable controller for all possible disturbances, but can be difficult to achieve and potentially yield poor performance [64]. Learning-based robust control improves performance by using data to learn a less conservative set and reduce the remaining model uncertainty. To guarantee safe robot control depends on the availability of prior knowledge and the types of uncertainties present. Research has demonstrated that machine learning methods, specifically reinforcement learning, can facilitate improved adaptability across broader categories of systems, requiring fewer pre-existing model assumptions [34]. In parallel, control theory, validation methods, statistical guarantees, and safety filters supply the valuable insights and structures essential for ensuring constraint satisfaction and stability guarantees throughout the learning process and in deployment [34].

Datasets and benchmarks play a significant role in the development of AI for AVs, especially in perception and end-to-end planning tasks. Key datasets include KITTI computer vision tasks on urban German roads, cityscapes, BDD100k, Mapillary Vistas data with segmentation masks, A*3D for collection scenes such as dark nights or rain or snow, automobile manufacturers datasets, and more. When data is needed, real-world testing is challenging, or testing is necessary at a fast and greater scale, simulators are used. For example, simulators such as CARLA, NVIDIA ISAAC, or AirSim might be used when training an AI AV. These simulators aim to replicate the real environment as closely as possible while still being computationally efficient to run faster than running in the real world, with a cheaper barrier to entry and no risk to the cars, infrastructure, or people involved. In between real-world data and simulations is using real-world data augmented into simulators and combining the two. As autonomous vehicles become more common, so will the data they collect. Some companies release this data to encourage development [65, 66]. However, transitioning from simulation to the real world can present major challenges [67].

5 Autonomous Vehicle Safety & Regulation

The regulation and policy landscape for autonomous vehicles has evolved significantly in recent years. Historically governing road rules internationally, the Vienna and Geneva Conventions on road traffic were reformed in 2016 to allow for automated vehicle features within ratified countries. In January 2021, an amendment proposal to the 1968 Vienna Convention was introduced, which was accepted by January 2022 and came into force in July 2022. In the United States, specific laws regarding autonomous vehicles have been enacted at the state level. In 2017, the SELF-DRIVE Act, aimed at expediting self-driving car adoption and restricting state-level performance standards, passed the House of Representatives but faced opposition in the Senate due to safety and liability concerns. In August 2022, bipartisan efforts were initiated in the House to reinvigorate self-driving vehicle legislation. The U.S. National Economic Council and Department of Transportation released the Federal Automated Vehicles Policy in September 2016, outlining standards for technology failures, passenger privacy, and accident scenarios to establish consistency and prevent a patchwork of state laws. The National Highway Traffic Safety Administration also issued documents in 2020 addressing occupant protection and safety standards for automated driving systems. California, a leader and hub for autonomous vehicle development, issues permits for testing and deployment on public roads. Nuro became the first manufacturer licensed to deploy fully autonomous cars in December 2020, followed by Cruise and Waymo in September 2021. Other states, such as Arizona, have allowed self-driving vehicles deployed in their cities. The technology to get a car on the pavement and transport people autonomously has arrived before thorough policy, regulatory, and standards development. Parallel to the still-developing AI technology powering the vehicles is a growing call for regulation and a scramble to control AI technology.

To encourage innovation and safety, the Autonomous Vehicles branch of the California DMV is working towards establishing autonomous vehicle testing and deployment guidelines and regulations. These cover permits, insurance, annual reports, reporting collisions and disengagements specifically, test driver requirements, and safety defects [68]. A recent white paper focused on autonomous vehicles surveys recent developments in automated technology and regulatory trends from a legal framing and prioritizes the analysis of liability and cybersecurity concerns for autonomous vehicles [7]. They conclude that a comprehensive AV-specific regulatory structure has not yet emerged at the federal or state level in the United States, and potential liability claims could include compliance with applicable laws or standards, quality control, warnings and instructions, and breach of warranties or false advertising. They suggest that companies will want to proactively show that their product makes/-made a reasonable decision. Autonomous vehicle customers, manufacturers, engineers, researchers, regulators, and policymakers will increasingly desire technical guarantees on the performance of mobile robots and the ability to explain the actions the robots take. Future issues for both policy and technical advancement include planning methods that provide safe and system-compliant performance in complex, uncertain, and cluttered environments, the direct propagation of uncertainty and features with safety guarantees, and verification and safety assessment methods that can be shared and trusted to bring autonomous vehicles safely and reliably into the real world and increase adoption [4, 7, 68].

As discussed previously, AI is an increasingly common approach in autonomous vehicle development [69]. Deep, fully convolutional neural networks as robust, flexible, high-capacity function approximators can model the complex relationship between sensory input and reward structure very well [4], and learning systems have been shown to perform well in collision avoidance [6]. AI models are known to be black boxes and challenging to explain. This problem must be solved for widespread autonomous vehicle adoption and well-informed regulation. The International Standard Organization Technical Committee 204, IEEE, and related standard organizations are prioritizing human well-being and assessing gaps in standardization for safe driving. Their standards demonstrate the need for explainability in AVs, covering human safety-related and data exchange standards, but no explainability standards currently exist [69]. Most work on explainable AI focuses on general AI, with specific models in narrow applications, whereas explaining autonomous vehicles will require explaining a large multi-goal-based and complex system with unique architectures and various interacting sub-systems [69, 70]. Explanations are a primary approach to building confidence and trust in autonomous vehicle technologies [70–72]. Many proposals for explaining AI call for an approach similar to the blackbox of an aircraft that can be recovered and analyzed after an incident. The National Transportation Safety Board calls for efficient event data recording in autonomous vehicles to provide plausible and faithful explanations in accident investigation [73]. NIST AI Risk Management Framework suggests that the safe operation of AI systems is improved through responsible design, development, and deployment practices, with clear information to stakeholders on the responsible use of the system, responsible decision-making by deployers and end users, and explanations and documentation of risks based on empirical evidence of incidents [74]. However, recovery-based approaches alone are not enough. Providing opportunities to proactively build trust and encourage collaboration with autonomous vehicle stakeholders (passengers, pedestrians, other road users, and regulatory boards) are necessary and useful in building trust in these systems [71, 72, 75–77]. It has been argued that trust is a substantial subjective predicting factor for adopting autonomous vehicles [78–80].

Explainability can help influence the acceptance of and reliance on autonomous systems [81]. For systems with humans in the loop, the NIST AI RMF recommends notification when a potential or actual adverse outcome caused by an AI system is detected. There exists work that outputs an explanation along with the result or action, but it can be difficult to train or craft details into the dataset before training [18]. Other work in explainable AI includes making existing or enhanced models interpretable, creating a second simpler-to-understand model that matches the deployed model, or working in natural language processing and computer vision for the generation of explanations derived from input [32, 33, 82–86]. The system's ability to perform safely might require breaking rules that call into question the ability to trust an autonomous system completely [87], and explanations in these rare cases will be especially important. The goal should not be to remove all errors or dynamic behavior from an autonomous system. For example, the system may be forced to take evasive or protective actions to remove themselves from a dangerous situation, to overtake an illegally parked car, or to pull over for an emergency vehicle. Not only will the planning system have to consider these when calculating and formulating the cost function, but enabling and explaining these actions may influence the overall trust in the system. Similar to the freezing robot problem [88], the robot freezes from an uncertainty explosion in future states that prevents any actions from being considered safe or possible. Calibrating the trust in automated vehicles will be an important feature of any application and must be considered for positive regulation [89]. Therefore, it will be a balance between trust, safety, and performance. Unfortunately, the existing regulations worldwide do not adequately address the explainability challenge facing autonomous vehicles [90].

Users of autonomous vehicles will spend time in the vehicle traveling from one destination to another. In a fully autonomous vehicle, the riders are free to work, have a conversation, eat, or watch entertainment. This produces ample data for the vehicle to track and process. With access to conversations, preferences, and location, data privacy in the autonomous vehicle context will require special consideration. Researchers have called for regulation such as the GDPR not to exempt vehicles and other robotic platforms as the world becomes increasingly connected and automated [69]. Additionally, the security of this data needs to be held to high standards. Article 12 of the GDPR on transparency requires providing information/explanation to data subjects to be intelligible, making it clear that future solutions in the autonomous vehicle space must be understandable and clear for stakeholders at various levels.

In both data and simulation, bias can be introduced to the system that must be considered when regulating autonomous vehicles. Proactive regulatory measures must consider how the autonomous vehicle is training and what data it is training on. Knowing this can give confidence that the autonomous vehicle will behave safely in the real world. In the future, verifying the performance and validity of their training methods without revealing trade secrets will be important to establishing mutually beneficial regulation.

A recent study indicated that for an autonomous car to be considered safe, it must have hundreds of millions of test miles, taking tens of years of driving time to complete [91]. Simulation and case-based testing are often used to check the performance of autonomous systems but alone do not provide the guarantees needed to ensure adequate safety and performance. Frameworks are called for that provide analytical and probabilistic proofs of safety as part of the simulation testing, rather than only checking the quantity of a finite set of traffic scenarios and simulations [4].

Autonomous mobile robots require interactive and cooperative decision-making in dangerous and dynamic environments. Other pedestrians, vehicles, and robot's intentions must be reasoned over and incorporated into the control planning system. These interactions often happen too fast or too often for direct communication to resolve any chance of collision. Therefore, a large body of work looks at minimizing the chance of collisions in planning trajectories of the other traffic participants and the system itself. Inherent uncertainties in the belief of the other agents and the environment make this a complex task. Verifying this prediction's safety helps verify the overall system's safety and is vital to any regulation of autonomous systems. Balance must be considered to mitigate risk yet allow the robot to continue operating even under high uncertainty to avoid the freezing robot problem [88]. We note that there are ways to tackle the freezing robot problem, including better descriptions of the environment and any obstacles, modeling the anticipated future information to reduce uncertainty, modeling cooperation based on a conditional formulation that models how the agents react to the robot's actions [92], or joint probability and cost distributions [93, 94]. However, it is not always possible to prevent the freezing-robot problem, even with perfect knowledge of all agent trajectories [93]. Many of these approaches make large assumptions about the knowledge or data available on the behaviors of other agents and may even require some level of control over the other agents, which is often not true.

Recent work surveying the future directions in autonomous vehicles calls for improvements to safe planning for imperfect sensor data, finding a balance of quality and speed, consistency in switching between different planners, and interpretability enhancement for learning-based planners [50]. Developing safe systems capable of operating in complex, cluttered environments while modeling the uncertain interaction with other traffic is an open and necessary challenge for mobile robots [4]. As future regulation is considered and crafted, we recommend the creators and auditors explore the abilities of safety, validation, and certification research in mobile robotics to provide more nuanced and well-informed regulation that builds a proactive and responsive approach to regulating mobile robots such as autonomous vehicles, and more generally in regulating AI.

6 A Safety Filter for Collision Avoidance

We present recent research that predicts other agents' behaviors and reactions during real-time execution to maintain social navigation norms and minimize risk. We work through one framework that utilizes probabilistic safety filters to enable robot deployment with more confidence. The framework produces confidence intervals that could be used in the given algorithm and could, in future work, be logged into a database, shared with human operators and auditors, visually demonstrated to pedestrians around the robot, or used as part of a larger system. These quantitative safety techniques can be advantageous for engineering purposes and can be accessed for human analysis, feedback, and regulation or control. We propose these methods as a foundation for discussing new regulatory tools. The original method is evaluated in a simulation, which provides insight into the system's expected performance and generates a discussion about using simulators and handling data in future regulatory frameworks.

We now consider learning a predictive safety filter for a given reinforcement learning policy that uses predicted agent trajectories, e.g., from long short-term memory networks [95, 96] or transformer architectures [97], along with uncertainty intervals around each prediction that are obtained using conformal prediction [98]. The result is an uncertainty-informed predictive safety filter that we can use as an example for more nuanced regulation [99]. This work uses conformal prediction, a distribution-free statistical tool [100–104], does not make any assumptions about the distribution of the agent trajectories, e.g., being Gaussian distributed, and showcases probabilistic safety guarantees for a learned controller in pedestrian-like environments. Approaches like this demonstrate the challenges of certifying a learning-based system and exemplify the complexities of guaranteeing safety, even when isolating one specific element of the overall autonomous vehicle framework. However, this shows the potential for metrics to guide regulation if regulatory parties require evidence of performance and safety standards and could help explain the reasoning behind a robot's actions. We explore the method's approach, assumptions, and results to show an example of safety guarantees in detail, covering an algorithm to train uncertainty-informed predictive safety filters for pre-trained reinforcement learning controllers using conformal prediction. The filter ensures probabilistic safety and incentivizes imitating the reinforcement learning policy. Evaluating the method's experiments in a widely-used reinforcement learning collision avoidance simulator [105] to consider certifying the safety of data, simulators, and deployment of learning-based systems.

6.1 Background

To plan in dynamic environments: Model Predictive Control (MPC) is the classic planning and control approach [106]. An MPC system could be a mobile robot navigating through a restaurant as it avoids collisions with the tables, chairs, customers, and staff. The robot selects a change in speed and heading (an action or series of actions) of minimum cost using predictions of the obstacles (static and dynamic) around them and the history, or state, of the environment they have seen so far. This is done in a receding horizon fashion, where the robot takes action and then receives new state updates from its sensor observations, and the process is repeated [107–109].

The robot, as an agent in the environment, uses the reactive-based methods to respond quickly to changes in the environment or other' agent's (such as the customers or staff or other robots) motion, with geometric or physics-based rules to ensure collision avoidance on each step [107–109]. These robots act in the moment, saving on computation cost and complexity–but risk producing sub-optimal trajectories. Reactive learning-based controllers are computationally efficient but often generate trajectories that may be inefficient, unnatural, and potentially dangerous [110–112].

In contrast, predictive-based methods first estimate the trajectories of other agents and then plan the system's actions by considering where the other agents will likely be. These methods yield a smoother plan but are more computationally expensive and may require additional knowledge about the other agents or risk incorrectly predicting the behavior of others, leading to dangerous situations. It is important to note that when predicting the behavior of other agents, overly conservative predictions can ultimately block the controlled robot's path, potentially slowing down or deadlocking all agents. This is known as the freezing robot problem [93]. The safest system may take no action. Therefore, it is important to balance safe planning and high performance.

Given expert demonstrations, such as data for the movement of staff and customers in the restaurant, some predictive and behavior-based planners can mimic what a human (or other robot) would do. These approaches estimate the cost functions of other agents, perform additional work to understand the intents of other agents , or mimic recorded trajectories[94, 113–118]. In recent years, learning-based controllers have increased in popularity to predict the trajectories of pedestrians, which can then be used in collision avoidance systems or evaluate the state of the environment and select the reaction-based next step. An example from recent work learns from complex and multi-modal distributions of agents' motions and predictions in real-time for planning [119]. Some work combines these approaches and integrates machine learning and model predictive control under uncertainty [120, 121].

Risk as uncertainty: There are many ways to consider risk in robotics [2, 8]. For our collision avoidance example, we define risk as the probability of collision with another robot or pedestrian. In a static environment with perfect sensor readings, the risk of collision would be low, if not zero. However, deploying a robot in the real world requires overcoming uncertainty. The behaviors of other agents and the quality of sensor readings may contain noise or error. The predictor used may produce inaccurate predictions when following a predictive-based collision avoidance method, i.e., uncertainty exists in the predictions of other agents.

In the safety filter of this work [99], conformal prediction (see [98]) is used for quantifying the uncertainty of trajectory predictions [122, 123]. Alternative methods model the underlying distribution as a Gaussian distribution and use Kalman filters or apply methods for finding safe sets such as forward and backward reachability [124–129]. These alternatives can be helpful but are often overly conservative, computationally complex, or make unrealistic assumptions. Similar to conformal prediction, a Bayesian framework provides probabilistic uncertainty quantification but requires access to prior knowledge about the distribution the data is sampled from, and probably approximately correct (PAC) learning theory can be used for producing upper bounds on the probability of error for a given algorithm and confidence level, but the results often involve large constants for the overall algorithmic error [130].

Conformal prediction provides distribution-free uncertainty quantification in such scenarios and has generally been used to quantify the uncertainty of machine learning models [98, 131–133]. Conformal prediction is an increasingly popular approach to obtain guarantees on a predictor's false negative rate, estimate reachable sets, and design model predictive controllers with safety guarantees [100–104]. The use of conformal prediction and similar methods is rapidly growing to guarantee and explain machine learning models. Recently, conformal prediction methods have been integrated with policy training for safety [134, 135], time series forecasting [123] and MPC of robots in dynamic environments [122, 136]. Another work combines conformal prediction with reachable sets for efficient and safe MPC [137]. These approaches integrate conformal prediction into an MPC, which reduces the framework's modularity and cannot be directly applied to learning-based controllers. In this example, we highlight how conformal prediction can be used broadly across machine learning applications and how it is used in the collision avoidance scenario as a filter to improve the safety of an existing controller. Even if a model could be trained to control the robot safely, the continuous updates and changes to the regulations and regulatory landscape necessitate the filtering and explainability of actions without requiring retraining the robot or starting anew.

Safety filters: Most safety techniques in reinforcement learning constrain the search during training updates, train with noise or adversarial agents, restrict the inputs to the policy, or attempt to learn the uncertainty of the entire system [34, 121, 138–140]. Work exists in the safe exploration of action spaces, such as filtering unsafe actions to prevent immediate negative consequences [141]. The training safety methods have been shown to perform negligibly better than non-safe reinforcement learning methods, and many robotic applications may prevent re-training, necessitating safety methods for pre-trained reinforcement learning controllers [141–144]. First introduced in [145], using safety filters for a controller in a closed-loop system has seen continued growth. Predictive safety filters assess if a proposed learning-based control input can lead to constraint violations and modify it if necessary to improve safety for future time steps. Other techniques exist, such as control barrier functions for verifying and enforcing system safety [127, 146–148], learning frameworks integrated with control barrier functions [149, 150], safety certification that continuously solves the optimization problem for a safe set at every online step [151], and model predictive control safety filters with system level synthesis have been proposed [152]. These approaches can provide system guarantees but require explicitly modeling a system's safety requirements, which are not trivial to design or implement, can be overly restrictive in the safe action sets, and increase online computational effort. We present recent work on a predictive safety filter for a pre-trained reinforcement learning policy to enable the use of high-performance, off-the-shelf controllers, increase generalizability to different applications, and provide stronger safety guarantees.

As the safety and risk criteria change, the safety filter's cost function can be updated in parallel without retraining the entire system. Modularity throughout the technical stack can improve the ability to regulate AI and provide third-party access to parts of the system without inspecting the entire framework and potentially risk losing trade secrets.

6.2 Safety Filtering Problem Formulation:

We first provide an overview of how the safety filtering problem was defined to produce a control policy (the safety filter) that closely follows a given policy (a pre-trained reinforcement learning controller) while ensuring that other dynamic agents are avoided [99]. The safety filter can be an additional tool to explain, understand, and regulate mobile robots. It can log information and be queried later to analyze the interactions or help explain the system's actions while also serving the original purpose of improving the system's safety. Let's consider the discrete-time dynamic control system:

$$x_{t+1} = f(x_t, u_t),\ x_0 := \zeta \tag{1}$$

Here, $x_t \in X \subseteq \mathbb{R}^N$ and $u_t \in U \subseteq \mathbb{R}^P$ denote the robot's state and the control input (or action chosen) at time $t \in \mathbb{N} \cup \{0\}$. The sets U and X denote the permissible actions and the system's workspace. The measurable function $f : \mathbb{R}^N \times \mathbb{R}^P \to \mathbb{R}^N$ describes the system dynamics and $\zeta \in \mathbb{R}^N$ is the initial condition of the system. The system operates in an environment with $A := \{1, 2, \cdots, m\}$ dynamic agents whose trajectories are unknown as they move from start to goal locations. This could be a vehicle moving down a street or through an intersection, with other agents such as cars, cyclists, and pedestrians moving around the ego vehicle. This representation is simple but exemplifies how most control problems are originally structured. Methods for auditing the problem formulation of embodied AI applications could be useful to help understand the system, its assumptions, and its weaknesses. This can be done without revealing how those deploying the robot solve the problem.

Let $\mathcal{D}$ be an unknown distribution over the other agent trajectories, and let $(\mathcal{T}_0, \mathcal{T}_1, \cdots) \sim \mathcal{D}$ describe a random trajectory where the stacked agent states $\mathcal{T}_t := (\mathcal{T}_t^1, \cdots, \mathcal{T}_t^m)$ at time t is drawn from $\mathbb{R}^{2m}$. In this example, each agent is assumed to operate in $\mathbb{R}^{2m}$ with a two-dimensional position, but the method is not limited to $\mathbb{R}^2$. The focus here is on the two-dimensional collision avoidance scenario, where all agents can be described by their position. We note that a fully autonomous vehicle will have a more complex system to consider but choose to highlight this critical element of path planning and the possibility for modularity in these systems and each subcomponent. As interdisciplinary work in this field grows, it will be important to consider how relevant high-level research is for helping guide real-world regulations and policies. While insights can still be gleaned from analyzing this system and the approach taken to improve safety, the system here is still overly simplified compared to real-world deployment. If not well understood, regulations taken directly from research could be misaligned with the practical systems they attempt to guide.

Let's use $\tau_t := \{\tau_t^1, \cdots, \tau_t^m\}$ when referring to a realization of $\mathcal{T}_t$ and assume access to the history of observations $\tau_{0:t} := \{\tau_0, \cdots, \tau_t\}$ online at time t. No assumptions are made on the form of the distribution $\mathcal{D}$ but do assume 1) that $\mathcal{D}$ is independent of the system (1), and 2) the availability of calibration data independently drawn from $\mathcal{D}$. We highlight here how the assumptions made when developing the system and safety filter are critical to understanding the risk of a system and how to verify safety. Disclosures of assumptions could improve transparency and help facilitate third-party testing or benchmarking against other techniques.

Assumption 1. *For any time $t \geq 0$, the control inputs $(u_0, \cdots, u_{t-1})$ and the resulting trajectory $(x_0, \cdots, x_t)$, following (1), do not change the distribution of $(\mathcal{T}_0, \mathcal{T}_1, \cdots) \sim D$.*

Assumption 1 holds approximately in many robotic applications, such as autonomous vehicles, where a car is likely to behave in ways that result in socially acceptable trajectories. The assumption is that the system is unlikely to drastically change the behavior of other agents, in which case conformal prediction still provides valid guarantees [153]. Later, distribution shifts are discussed in more detail and how they affect the possibility of certifying a system such as this as safe. The second assumption assumes the availability of training and calibration data drawn from $\mathcal{D}$.

Assumption 2. *Access to a dataset of trajectories* $D := \{\tau^{(0)}, \tau^{(1)}, \cdots, \tau^{(K)}\}$ *in which each of the* K *trajectories* $\tau^{(i)} := \{\tau_0^{(i)}, \tau_1^{(i)}, \cdots\}$ *is independently drawn from* $\mathcal{D}$*.*

With the ability to collect large amounts of data from rapidly advancing high-fidelity simulators or robotic applications such as autonomous vehicles where datasets are becoming widely available, assumption 2 is not considered restrictive. Lastly, the dataset D is split into training datasets D_{Ytrain} and D_{train} from which the trajectory predictor and safety filter will train, respectively, and a calibration dataset D_{cal} to quantify the uncertainty of the trajectory predictor. Given, collected, or curated benchmark data could be used to test the filters and models to ensure safety. In autonomous vehicles, benchmark datasets could be useful tools when providing permits and assessing performance, just as human drivers must also pass tests to be licensed.

The assumptions engineers and researchers make are critical in defining the applicability of a given solution. The assumptions made for this example are idealized. However, they hold approximately in practice if the controlled ego agent takes conservative actions that have little influence on the behavior of other agents. This is the case in many robotic applications, such as autonomous vehicles, where a car is likely to behave in ways that result in socially acceptable trajectories. This may be an acceptable assumption in research, but when deploying in the real world, this assumption should be disclosed to the stakeholders. Parallels could be drawn in future work to the disclosure of how data is processed and handled. These assumptions may give some insight into the solutions but do not give the solution away. Assessment of assumptions is vital for verifying the safety of autonomous robots. The algorithm uses uncertainty quantification for in-distribution data, meaning that the distribution $\mathcal{D}$ during training should be the same as during testing. Interactions, i.e., couplings between x and $\mathcal{D}$, may lead to a test distribution different from the training distribution, resulting in a distribution shift. A small distribution shift will not significantly affect the algorithm or guarantees, as supported by Corollary 2.1, where it is formally shown that small distribution shifts only lead to small deviations from the desired conformal prediction guarantees [153].

Pedestrian interactions happen when the agents are nearby and force an agent to change their path. In the experiments, it is unlikely that this will drastically change the behavior of other agents, i.e., we do not expect a large distribution shift. As done in the paper, this can be verified empirically by plotting the distribution of the distances between agents. Therefore, it can be reasonably assumed that the interactions are not causing a significant shift in the agent trajectories. Ensuring accurate and ample data collection is also necessary before deployment to deal with such distribution shifts in practice. More data does not necessarily mean a better solution, but the data used during training will directly influence the system's performance. To regulate an AI system should also weigh heavily the regulation of the data that the

AI system trained and tested on. For this example, if more protection against distribution shifts is desired, implementing other conformal prediction techniques, such as adaptive conformal prediction [136, 154], may be necessary. However, the theoretical guarantees in most adaptive conformal prediction settings are weaker. We call for regulators and policymakers to consider this nuanced version of AI, where subtle algorithm changes can drastically change the system and how it should be regulated.

Given a pre-defined controller $\pi : \mathbb{R}^N \times \mathbb{R}^{2m} \to \mathbb{R}^P$ providing the control inputs

$$u_t := \pi(x_t, \tau_t). \tag{2}$$

Here, the policy π is an reinforcement learning policy that aims to reach a final location while avoiding collisions with agents in A. However, the safety filter may take dangerous actions if deployed as is. With no additional knowledge of how it will behave or if risky actions will be avoided, the robot could collide with the other agents at a rate unacceptable by stakeholders [99].

6.3 Conformal Predictive Safety Filter

The goal is to add a safety filter $\hat{\pi}$ to the pre-defined reinforcement learning policy π that may not have any safety certification or is only valid under certain assumptions, e.g., $\mathcal{D}$ being Gaussian to demonstrate a technique for certifying the safety of an reinforcement learning model in embodied AI. Here, a trajectory predictor Y is learned to predict future agent trajectories from past agent observations and conformal prediction to obtain uncertainty intervals for these predictions. The predictive safety filter uses this information to achieve the safety of the reinforcement learning policy π while minimally deviating from π [99]. We briefly explain the offline training of the safety filter $\hat{\pi}$, the results presented, and the implications for regulation.

Trajectory predictor: Given a prediction horizon H and the history of agent observations $\tau_{0:t}$, let's define a trajectory predictor $Y : \mathbb{R}^{(t+1)2m} \to \mathbb{R}^{2mH}$ that predicts the H future agent states $(\mathcal{T}_{t+1}, \ldots, \mathcal{T}_{t+H})$ as $\bar{\tau}_{t+1:H} := Y(\tau_{0:t})$ where

$$Y(\tau_{0:t}) := (\bar{\tau}_{t+1}, \cdots, \bar{\tau}_{t+H}). \tag{3}$$

For training Y, we independently sample from $\mathcal{D}$ a dataset $D_{Y train}$ with trajectories from time 0 to time T, i.e., $\tau_{0:T}^{(i)} := (\tau_0^{(i)}, \cdots, \tau_t^{(i)}, \tau_{t+1}^{(i)}, \cdots, \tau_T^{(i)})$ is the ith trajectory in the dataset $D_{Y train}$. We could use any trajectory predictor Y, e.g., long short-term memory networks [95, 96] or transformer architectures [97]. In complex autonomous vehicles or other mobile robots and applications of embodied AI, multiple predictors and sensors could be used, requiring multiple training datasets, models, and simulators to consider for regulatory purposes. Each dataset, method of learning, and simulator will require unique assumptions and pose diverse challenges for ensuring safety, accuracy, and fairness. In this example, we will consider training a long short-term memory

(LSTM) network by minimizing over:

$$\min_{Y} \frac{1}{|D_{Ytrain}|} \sum_{i=1}^{|D_{Ytrain}|} \|\tau_{t+1:H}^{(i)} - Y(\tau_{0:t}^{(i)})\|^2 \tag{4}$$

Adding safety with confidence intervals: Conformal prediction is used to construct regions around the predicted trajectories that contain the true but unknown trajectory with high probability. This statistical tool increases the system's confidence around the uncertainty. A potentially useful tool for regulating specific components of complex modular systems. We refer the reader to [98] for a general introduction to conformal prediction. We briefly summarize [99, 122, 123], where the authors present a technique to construct valid prediction regions applied to recurrent neural networks. Given observations $\tau_{0:t} := (\tau_0, \cdots, \tau_t)$ at time t, where $\tau_t := (\tau_t^0, \cdots, \tau_t^m)$, we can use the trajectory predictor Y to obtain predictions $\overline{\tau}_{t+1:H} := (\overline{\tau}_{t+1}, \cdots, \overline{\tau}_{t+H})$ for the specified prediction horizon H. Given a failure probability of $\delta \in (0, 1)$, we seek values $C_{t+1:H} := (C_{t+1}, \cdots, C_{t+H})$ as prediction intervals around each prediction such that:

$$\text{Prob}(||\tau_{t+h} - \overline{\tau}_{t+h}|| \leq C_{t+h},\ \forall h \in \{1, \cdots, H\}) \geq 1 - \delta \tag{5}$$

First, define the non-conformity score function $R_{t+h} := ||\tau_{t+h} - \overline{\tau}_{t+h}||$ and evaluate it across a conformal calibration dataset D_{cal} that is independently sampled from the distribution $\mathcal{D}$ defined previously. Here, a small non-conformity score corresponds to accurate predictions, and a large score indicates a more inaccurate Y. Second, sort the non-conformity scores from the calibration dataset D_{cal} in non-decreasing order and append infinity as the $(|D_{cal}|+1)$-th value. Third, define $p := \lceil(|D_{cal}|+1)(1-\bar{\delta})\rceil$ and let the prediction interval C_{t+h} correspond to the pth quantile over the sorted non-conformity scores for each prediction step $h \in \{1, \ldots, H\}$. From [99, 122, Theorem 1], set $\bar{\delta} := \delta/T$ to ensure collision avoidance with a probability of at least $1 - \delta$ across the steps. Finally, when making predictions online, use the values $C_{t+1:H}$ as prediction intervals around each predicted trajectory $\overline{\tau}_{t+1:H}$ as done in [99]. These C values provide insight into the uncertainty in the predictions and, thus, the uncertainty in our controller. Logging, analyzing, and visualizing these values could be beneficial for assessing the performance of individual elements of a modular system and setting standards for each component. It also provides an opportunity for aiding human-robot interactions through visual explainability, demonstrating the belief the system holds.

Training: To train the predictive safety filter $\hat{\pi}$, 1) forward simulate the system from (1) under the nominal reinforcement learning policy π from (2) using the trajectory predictions from (3), and 2) enforce that the trajectory of the system from (1) under the safety filter $\hat{\pi}$ imitates the trajectory under the reinforcement learning policy π while incorporating the conformal predictions regions $C_{t+1:H}$ to account for uncertainty in the trajectory predictions.

For the first step, use the pre-defined reinforcement learning policy π and the predictions from Y to simulate the system dynamics under the reinforcement learning policy forward into the future by H. From this, obtain the nominal future trajectory

$\bar{x}_{t+1:H} := (\bar{x}_{t+1}, ..., \bar{x}_{t+H})$ as

$$\begin{aligned}
\bar{x}_t :=& x_t \\
\bar{u}_t :=& \pi(x_t, \tau_t) \\
\bar{x}_{t+h} :=& f(\bar{x}_{t+h-1}, \bar{u}_{t+h-1}), \, \forall h \in \{1, ..., H\} \\
\bar{u}_{t+h} :=& \pi(\bar{x}_{t+h}, \bar{\tau}_{t+h}), \, \forall h \in \{1, ..., H-1\}
\end{aligned} \tag{6}$$

The safety filter $\hat{\pi} : \mathbb{R}^{2mH} \times \mathbb{R}^{HM} \times \mathbb{R}^{HN} \times \mathbb{R}^{H} \to \mathbb{R}^{HP}$ optimizes the following objective:

$$\begin{aligned}
&\min_{\hat{\pi}} \sum_{h=1}^{H} ||\bar{x}_{t+h} - \hat{x}_{t+h}||^2 \\
\text{s.t.} \quad & \hat{x}_t := x_t \\
& \hat{x}_{t+h} := f(\hat{x}_{t+h-1}, \hat{u}_{t:H-1}(h)), \, \forall h \in \{1, ..., H\} \\
& \hat{u}_{t:H-1} := \hat{\pi}(\bar{\tau}_{t+1:H}, \bar{u}_{t:H-1}, \bar{x}_{t+1:H}, C_{t+1:H}) \\
& ||\bar{\tau}^j_{t+h} - \hat{x}_{t+h}|| \geq C_{t+h} + \epsilon, \, \forall h \in \{1, ..., H\}, \forall j \in A
\end{aligned} \tag{7}$$

The safety filter produces $\hat{u}_{t:H-1} := \hat{\pi}(\bar{\tau}_{t+1:H}, \bar{u}_{t:H-1}, \bar{x}_{t+1:H}, C_{t+1:H})$, where $\hat{u}_{t:H-1}$ contains control inputs for the next H time steps. $\hat{u}_{t:H-1}(h)$ refers to accessing the control input for the hth time step. The filter is recursively applied to the system across timesteps T. Intuitively, the safety filter minimizes the distance between the nominal reinforcement learning trajectory $\bar{x}_{t+1:H}$ and the safety filter trajectory $\hat{x}_{t+1:H}$, i.e., the trajectory obtained under the safety filter $\hat{\pi}$. Additionally, the safety filter trajectory $\hat{x}_{t+1:H}$ should avoid the agent predictions $\bar{\tau}^j_{t+h}$ (for all agents $j \in A$) and uncertainty intervals C_{t+h}. Specifically, enforcing the safety constraint $||\tau^j_{t+h} - \hat{x}_{t+h}|| \geq \epsilon$, where τ^j_{t+h} is unknown, and $\epsilon > 0$ is a user-defined minimum collision avoidance distance, where $||\bar{\tau}^j_{t+h} - \hat{x}_{t+h}|| \geq C_{t+h} + \epsilon$ holds. Since we know $\text{Prob}(||\tau_{t+h} - \bar{\tau}_{t+h}|| \leq C_{t+h}, \, \forall h \in \{1, \ldots, H\}) \geq 1 - \delta$, it is ensured that $\text{Prob}(||\tau^j_{t+h} - \hat{x}_{t+h}|| \geq \epsilon, \, \forall h \in \{1, \ldots, H\}) \geq 1 - \delta$. Sampling another dataset D_{train} of independent trajectories from $\mathcal{D}$ to train the safety filter. Combine each training trajectory's corresponding predictions and nominal reinforcement learning trajectories into the training dataset $D_{sftrain}$ to solve (7). For example, train a multi-layer feedforward neural network $\hat{\pi}$. This solves (7) approximately over the training set $D_{sftrain}$ and rewrites the constrained optimization problem (7) into an unconstrained optimization problem. Therefore, the original cost function is augmented with the constraints of (7) that are minimized until convergence.

Online execution: During runtime the safety filter can be used by first computing the set of H next control inputs $\hat{u}_{t:H-1} := \hat{\pi}(\bar{\tau}_{t+1:H}, \bar{u}_{t:H-1}, \bar{x}_{t+1:H}, C_{t+1:H})$ in a receding horizon manner by applying the first element $\hat{u}_{t:H-1}(1)$ at each time step t. Then, pass the history of trajectories $\tau_{0:t}$ to the predictor Y to obtain predictions $\bar{\tau}_{t+1:H}$. After, we obtain the nominal reinforcement learning control inputs and trajectory $\bar{u}_{t:H-1}$ and $\bar{x}_{t+1:H}$, as well as the conformal uncertainty intervals $C_{t+1:H}$. Finally, compute the safety filtered control input and apply $\hat{u}_{t:H-1}(1)$ to the system. The quantified uncertainty intervals can then be logged, inspected, and analyzed, as well as the history of actions taken and predictions made. These represent the collision avoidance and motion control system's beliefs. We refer the reader to the safety filter work for a full description of the algorithm, experiments, and results [99].

Guarantees: Under the assumption that the safety filter achieves the objective in (7), the safety filter guarantees probabilistic safety, i.e., that $\text{Prob}(||\tau_t^j - \hat{x}_t|| \geq \epsilon,\ \forall t \in \{1,\ldots,H\}) \geq 1 - \delta$ as we use $\bar{\delta} := \delta/T$ [99]. For uncertainty quantification of the predictor, δ can be tuned to produce larger avoidance actions or smaller ones, depending on the level of conservatism desired. Note that these guarantees hold under idealized assumptions. In practice, there may be reasons why a safety system does not achieve exact probabilistic coverage. One reason is the approximate solving of the optimization problem (7) over the training set $D_{sftrain}$ by rewriting (7) as an unconstrained optimization problem. In the experimental results, however, it is shown that the safety filter $\hat{\pi}$ trained with sufficient data performs well in practice and that the method provides an added layer of safety over the baseline reinforcement learning policy. Any standards or regulations should consider the assumptions made and the optimization problem solved. Future regulation should pursue diving deeply into different architectures and algorithms to better fit regulation to the systems being deployed.

It is important to note that the prediction intervals C_{t+h} naturally depend on the underlying distribution $\mathcal{D}$, the predictor's accuracy, and the user-specified risk tolerance δ. Without the intervals C_{t+h}, the safety filter $\hat{\pi}$ would only mimic the baseline reinforcement learning controller π. Therefore, the safety filter $\hat{\pi}$ may perform differently than π. This attribute enables potentially different policies with the learned safety filter and variable levels of desired safety to be enforced. When discussing regulation at this level more flexibility can be written in to allow innovation while quantifiably assessing and governing autonomous systems–without requiring perfection.

As per Assumption 1, it assumes that the distribution of trajectories does not change from offline training to online testing. Small distribution shifts in $\mathcal{D}$ will not significantly affect the algorithm and its guarantees. A supporting argument of this claim is given in [153], where it is formally shown that small distribution shifts only lead to small deviations from the desired conformal prediction guarantees. In the experiments, it is explicitly checked that agent interaction does not introduce a larger distribution shift. One may increase the robustness of the prediction intervals using adaptive conformal prediction [136] or other methods. For each safety feature implemented, it too must be analyzed for possible errors and risks. Here, if the assumptions are broken, then no safety is guaranteed. Requiring companies to produce assumptions and validate them could help support certifying embodied AI.

6.4 Experimental Evaluation

The experiments in [99] were run in the Collision Avoidance Gym [105]. The multiagent gym environment is open-sourced and provides pre-trained reinforcement learning policies for navigation from given start to goal locations. Agents can sense the locations and velocities of the other agents. Simulators like this enable rapid production and testing, and the gym allows customizable dynamic models and supports quick benchmark testing against future techniques. The authors demonstrate the transferability of policies from training in their gym simulator to deployment on real-world aerial and ground robots in real-time with diverse sensors [105].

The pre-trained reinforcement learning controller without the safety filter performed less conservative trajectories, leading to a shorter time to goal but a higher number of collisions than a more conservative method. However, the conservative method performed overly conservative trajectories that reduced collisions but increased failures to reach the goal. The conformal predictive safety filter with $\delta := 0.01$ minimized the negative side effects of each control policy, with fewer collisions and failures. The results matched the risk tolerance set by $\delta := 0.01$, with the nonzero collisions under a configurable threshold. GA3C, without the safety filter, had the shortest distance from other agents and time to goal, but the safety filter performs as well as the other algorithms in time to goal and minimally impacts GA3C. Suggesting that safety filters can be applied with minimal impact on performance. It is noted that the simulated gym environment is dense, and the distances between agents can be small, with subtle changes during key interactions. Fine-tuning the hyperparameters or δ could produce more drastic changes as desired and should be analyzed when assessing the system's safety or adherence to standards and regulations. The learned safety filter algorithms that mimic the original policies may differ from the expert policy due to conformal prediction constraints during learning, imperfect training, and finite training data. Overall, the safety filter algorithms outperform the non-safety filter algorithms in the number of collisions and failures. The safety filter algorithms maintain similar time to goals and distances from other agents and show an awareness of the uncertainty in the underlying approaches.

7 Regulating Under Uncertainty

Subsequently, we discuss the potential use of certification and probabilistic guarantees as new risk regulation tools not currently used and the potential to contribute to public and stakeholder assessment, input, performance upkeep, and accountability mechanisms. We relate these ideas to current conversations of AI regulation, for example, the recent bill of rights published by the White House, reports from the FTC and the Artificial Intelligence Commission advising AI companies and government parties, the GDPR's right to explainability, and other scholarly works. However, in these recent works, AI is considered broadly and vaguely. The five principles from the White House's blueprint can be consulted here, with more specific context using our presented example. Doing so highlights how difficult regulating AI will be but demonstrates that tools exist to regulate it in specific and well-informed ways. Using research in safety guarantees specific to each application in regulation could help foster public acceptance, performance accountability, third-party assessment, and adaptive regulation.

The first principle for protection from unsafe or ineffective systems calls for diverse communities, stakeholders, and experts to be consulted and part of the team when developing the systems. A system should undergo pre-deployment testing, risk identification, and mitigation with ongoing monitoring. In our example, the addition of a safety filter led to an added discussion around the system and the communities involved in developing it and the original reinforcement learning controller. Adding filters and other safety techniques brings new ideas and results to help explain, monitor, and mitigate the problems the original system might encounter. This helped us develop a system with safe goals that does not endanger stakeholders, can be evaluated independently, and does not unreasonably hinder performance. The conformal prediction technique enables operators or auditors to set and assess the acceptable risk level in the predictions. This forces developers to set a level of risk and allows other stakeholders to voice concerns, argue for higher standards, and monitor the performance concerning the intended safety level.

Regulators should utilize simulations as benchmarks due to their capacity to realistically assess risk, cost-effectively test policies, facilitate data-driven decision-making, conduct scenario analysis, respond swiftly to changing environments, engage stakeholders effectively, aid in training and education, prevent unintended consequences, support continuous improvement, and address ethical concerns. A simulator's fidelity and ability to accurately train and test a system's performance are crucial to safe deployments. The gym environment used in the example experiments is too simple to certify the system as safe for all autonomous vehicle operations in the real world but shows experimental evidence for safety in pedestrian robot navigation [6]. Regulators can set standards for learning-based systems, add comparison capabilities, and build adaptable regulatory frameworks by leveraging simulations as benchmarks and certifying the quality of simulators.

The second principle asserts that people should not face discrimination by algorithms, and systems should be used and designed equitably. Similar to the distribution shift analysis, implementing the safety filter led to a discussion about how this system will interact with the other agents and how it will treat agents that behave differently. The predictor's performance can drop significantly when predicting agent behaviors that do not perform similarly to the training and calibration data. This highlights the need to assess the datasets used in training for bias towards certain behaviors or interactions in our example. We found minimal distribution shift in our simplified simulations and can move forward in confidence, but we would have had to collect a higher quality or amount of data if the distribution shift was significant. Algorithms that can affect or use a protected class (and its proxies) for reasoning should be proactively designed to counteract the potential for bias, discrimination, and distribution shifts. This analysis can be done without revealing trade secrets or private information. Explainability techniques should be considered for the model and the dataset and simulations used in training. Implementing a safety filter helps inspire these considerations and hopefully limits any negative outcomes of bias and discrimination.

The third principle calls for data privacy and that users should be protected from privacy violations through design choices that ensure protections by default and provide the user agency over how their data is used or sold. Here, we must consider how the data was collected and handled in our example. We collected large amounts of trajectory data, possibly leading to data security and privacy issues if handled incorrectly. We propose logging and analyzing additional data, which must be handled carefully to protect all stakeholders and follow the already established and growing data privacy regulations. Most of the safety and validation techniques mentioned in this paper produce additional data requiring additional security and privacy work.

The fourth principle suggests that automated systems must be understandable to the users they impact. The same confidence intervals could be turned into explanations for how the robot behaved and how the robot will behave in interactions with pedestrians around the robot and those operating the robot. Automated systems should provide technically valid, meaningful, and useful explanations–calibrated to the level of risk based on the context. We suggest reporting on these confidence intervals, how accurate they are, and using them to explain instances that appear outside of the intervals or cause a distribution shift. This could include summaries of information in plain language and clarity of assessments provided to the public whenever possible.

The final principle states that human alternatives and the ability to opt-out should be possible where appropriate, with the possibility of appealing an automated decision, especially in high-risk settings. If an autonomous vehicle can display its reasoning and beliefs, such as the confidence intervals, it would allow its passengers to feel more confident in the vehicle or remove themselves from the service. In the event of a crash, the safety filter data could provide valuable insight into how the robot made its decision and if the idea of statistical reasoning extends beyond this one subcomponent could help settle disputes around liability.

Diving deeper: Although there have been significant strides in safety for AI and robotics research, ensuring safe robot interactions remains an open problem [155]. It is important to consider all the work that has been done and can be used in regulating AI that already exists, and at the same time, we stress the importance of building in flexibility to adapt as new techniques are developed to address the ever-changing technological landscape. From the collision avoidance example, we highlight that focusing on a specific application and framework allowed statistical guarantees to increase safety for the robotic system. The same technique could be used to inform the regulatory structures and supervisory enforcement mentioned previously in the background of AI regulation. To do so will require more open interfacing between companies and regulatory oversight to build checks and balances for specific applications. The field of AI is incredibly broad, and generalizing one solution for another could stifle progress and innovation. We note that the application of the safety filter here was tuneable to fit the desired safety levels, generalizable to predictive machine learning models, and did not hinder the robot's performance.

The assumptions engineers and researchers make for each application are critical in defining the safety of a given solution. The assumptions made for this example were idealized. However, they held approximately in practice, assuming the actions taken had little influence on the behavior of other agents. This may be an acceptable assumption in research, but when deploying in the real world, we argued that this assumption should be disclosed to the stakeholders. These assumptions can yield insight into the solutions and do not give the solution away. Assessment of assumptions is vital for verifying the safety of autonomous robots. Parallels could be drawn to the disclosure of how data is processed and handled. The algorithm used uncertainty quantification for in-distribution data, meaning that the distribution $\mathcal{D}$ during training should be the same as during testing. These couplings between x and $\mathcal{D}$ could have led to the test distribution being different from the training distribution, causing a dangerous distribution shift. It is important to assess the assumptions made and guarantee their validity, as done in the example. This can only be done when considering the specifics of a given application and solution. Finding ways to standardize the auditing of assumptions could be beneficial for regulating AI.

Modular solutions: For engineers developing these complex systems, building one end-to-end system that solves the entire problem is attractive. However, most autonomous systems deployed today are modular, and this trait should be taken advantage of to regulate the subcomponents of a system. In our example, we provide probabilistic guarantees on the system and stronger guarantees on the predictor. Combining this technique with other validation methods could strengthen our guarantees of the overall system. In a more complex model of x we may have to consider other techniques and more subsystems to validate. To certify the safety of an autonomous vehicle will require separate techniques for localization, static object detection, dynamic object detection, scene understanding, tracking, prediction, planning, control, simulation technologies, testing frameworks, and human-machine interactions. Systems must be in place to spot liability issues in each subcomponent, which separate parties may manufacture and develop. The problems introduced by AI solutions can appear differently in each category and require a nuanced approach considering how safety, fairness, and reliability should be considered for each application. For our example, the safety for our 2-dimensional interactions was easy to understand, define, and explain. However, in a more complicated real-world autonomous vehicle scenario, more complex issues such as bias, privacy, security, vehicle safety, liability prioritization, and more arise. This will require an interdisciplinary approach.

Quantifying uncertainty: When considering safety and risk, the most important aspect of robotics is quantifying the uncertainty. From the initial definition of a problem to the widespread deployment and use of the system, the uncertainty should be defined, listed clearly, and quantified to monitor and verify reasonable actions under uncertainty. If a system cannot be proven safe theoretically, it should verify a strong understanding of the uncertainty and provide self-regulation methods to check these assumptions and how they evolve over deployment. Appropriate measures should be taken when the uncertainty level is too high and degrades performance or risks the safety of any stakeholders. The NIST AI RMF recommends notification when a potential or actual adverse outcome caused by an AI system is detected, which could be achieved by using the confidence intervals to flag interactions that contain high uncertainty to a desired δ level. In our example, the C values provide insight into the uncertainty in the predictions and, thus, the uncertainty in our controller. The National Transportation Safety Board calls for efficient event data recording in autonomous vehicles to provide plausible and faithful explanations in accident investigations. Logging, analyzing, and visualizing these C values could be beneficial for assessing the performance of individual elements of our modular system and setting standards for each component. They also provide an opportunity for aiding human-robot interactions through visual explainability, demonstrating the belief our system holds. Additionally, as the DoD's TEVV suggests, controlling δ allows us to confidently deploy a system without requiring perfection.

8 Conclusion

We use an example of adding a safety filter to a collision avoidance application as the initial step towards considering safety, validating assumptions, and quantifying uncertainty to improve the overall system performance and societal interactions with robots. We call for work in regulating AI to similarly consider diving deeply into specific applications and using new techniques in explaining, filtering, validating, and assessing AI performance when drafting rules, standards, or regulations in a nuanced and well-informed approach. Being well-informed and specific will require more interdisciplinary work but will enable balancing innovation and protection.

In conclusion, we leverage recent research in reinforcement learning safety to ignite inclusive and technical discussions around positive opportunities that balance robotic regulation and human safety without restricting innovation. We hope work in this area will increase public acceptance, performance accountability, third-party assessment, and adaptive regulation. We invite further dialogue from interdisciplinary researchers across robotics, computer science, law, and ethics that builds from these tools. We aim to spark informed regulation for bettering these robotic systems, those who develop them, and those they affect.

References

[1] Li, F.-F., Reich, R.: Stanford hai faculty urge president biden to approach ai with a moonshot mentality. Stanford HAI' (2023)

[2] Kaminski, M.E.: Regulating the risks of ai. Forthcoming, Boston University Law Review **103** (2023)

[3] Isele, D., Nakhaei, A., Fujimura, K.: Safe reinforcement learning on autonomous vehicles. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1–6 (2018). IEEE

[4] Schwarting, W., Alonso-Mora, J., Rus, D.: Planning and decision-making for autonomous vehicles. Annual Review of Control, Robotics, and Autonomous Systems **1**, 187–210 (2018)

[5] Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable ai systems. ACM Transactions on Interactive Intelligent Systems (TiiS) **11**(3-4), 1–45 (2021)

[6] Everett, M., Chen, Y.F., How, J.P.: Collision avoidance in pedestrian-rich environments with deep reinforcement learning. IEEE Access **9**, 10357–10377 (2021)

[7] Autonomous vehicles: Legal and regulatory developments in the united states (2021)

[8] Thomasen, K.: Safety in artificial intelligence & robotics governance in canada. Available at SSRN (2023)

[9] Willis, L.E.: Performance-based consumer law. The University of Chicago Law Review, 1309–1409 (2015)

[10] Allen, H.J.: Regulatory sandboxes. Geo. Wash. L. Rev. **87**, 579 (2019)

[11] Anderljung, M., Barnhart, J., Leung, J., Korinek, A., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., et al.: Frontier ai regulation: Managing emerging risks to public safety. arXiv preprint arXiv:2307.03718 (2023)

[12] Commission, F.T., et al.: Combatting online harms through innovation. Technical report, Tech. rep. June 16, 2022.: https://www. ftc. gov/reports/combattingonline . . . (2022)

[13] Hine, E., Floridi, L.: The blueprint for an ai bill of rights: in search of enaction, at risk of inaction. Minds and Machines, 1–8 (2023)

[14] Sanchez-Graells, A.: Governing the assessment and taking of risks in digital procurement governance. To be included in A Sanchez-Graells, Digital Technologies and Public Procurement. Gatekeeping and experimentation in digital public governance (OUP, forthcoming) (2022)

[15] Algorithmic accountability act of 202. 117th U.S. Congress (2022)

[16] Clark, M., Kearns, K., Overholt, J., Gross, K., Barthelemy, B., Reed, C.: Air force research laboratory test and evaluation, verification and validation of autonomous systems challenge exploration. Air Force Research Lab, Wright-Patterson AFB, Tech. Rep. (2014)

[17] Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI magazine **38**(3), 50–57 (2017)

[18] Hind, M., Wei, D., Campbell, M., Codella, N.C., Dhurandhar, A., Mojsilović, A., Natesan Ramamurthy, K., Varshney, K.R.: Ted: Teaching ai to explain its decisions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 123–129 (2019)

[19] Diakopoulos, N.: Algorithmic accountability reporting: On the investigation of black boxes (2014)

[20] Sweeney, L.: Discrimination in online ad delivery. Communications of the ACM **56**(5), 44–54 (2013)

[21] Lim, B.Y., Dey, A.K.: Assessing demand for intelligibility in context-aware applications. In: Proceedings of the 11th International Conference on Ubiquitous Computing, pp. 195–204 (2009)

[22] Diakopoulos, N.: Enabling accountability of algorithmic media: transparency as a constructive and critical lens. Transparent data mining for Big and Small Data, 25–43 (2017)

[23] Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., et al.: Accountability of ai under the law: The role of explanation. arXiv preprint arXiv:1711.01134 (2017)

[24] Mittelstadt, B.: Automation, algorithms, and politics— auditing for transparency in content personalization systems. International Journal of Communication **10**, 12 (2016)

[25] Turilli, M., Floridi, L.: The ethics of information transparency. Ethics and Information Technology **11**, 105–112 (2009)

[26] Jones, G.W.: The search for local accountability. Strengthening local government in the 1990s, 49–78 (1992)

[27] Martinho, A., Herber, N., Kroesen, M., Chorus, C.: Ethical issues in focus by the autonomous vehicles industry. Transport reviews **41**(5), 556–577 (2021)

[28] Collingwood, L.: Privacy implications and liability issues of autonomous vehicles. Information & Communications Technology Law **26**(1), 32–45 (2017)

[29] Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M.A., Liang, P.: Foundation models and fair use. arXiv preprint arXiv:2303.15715 (2023)

[30] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., et al.: The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228 (2018)

[31] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z.: Xai—explainable artificial intelligence. Science robotics **4**(37), 7120 (2019)

[32] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

[33] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)

[34] Brunke, L., Greeff, M., Hall, A.W., Yuan, Z., Zhou, S., Panerati, J., Schoellig,

A.P.: Safe learning in robotics: From learning-based control to safe reinforcement learning. Annual Review of Control, Robotics, and Autonomous Systems **5**, 411–444 (2022)

[35] Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C.: Auditing algorithms: Research methods for detecting discrimination on internet platforms. Data and discrimination: converting critical concerns into productive inquiry **22**(2014), 4349–4357 (2014)

[36] Faisal, A., Kamruzzaman, M., Yigitcanlar, T., Currie, G.: Understanding autonomous vehicles. Journal of transport and land use **12**(1), 45–72 (2019)

[37] Scott, J., Scott, C.: Drone delivery models for healthcare (2017)

[38] Forlizzi, J., DiSalvo, C.: Service robots in the domestic environment: a study of the roomba vacuum in the home. In: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction, pp. 258–265 (2006)

[39] Transportation Federal Highway Administration, U.S.D.: U.s. driving increases for sixth straight year, new federal data show (2017)

[40] National Highway Traffic Safety Administration, U.S.D.o.T.: Traffic safety facts, crash, stats (2022)

[41] Kuutti, S., Bowden, R., Jin, Y., Barber, P., Fallah, S.: A survey of deep learning applications to autonomous vehicle control. IEEE Transactions on Intelligent Transportation Systems **22**(2), 712–733 (2020)

[42] Alonso-Mora, J., Samaranayake, S., Wallar, A., Frazzoli, E., Rus, D.: On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. Proceedings of the National Academy of Sciences **114**(3), 462–467 (2017)

[43] Alonso-Mora, J., Wallar, A., Rus, D.: Predictive routing for autonomous mobility-on-demand systems with ride-sharing. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3583–3590 (2017). IEEE

[44] Luettel, T., Himmelsbach, M., Wuensche, H.-J.: Autonomous ground vehicles—concepts and a path to the future. Proceedings of the IEEE **100**(Special Centennial Issue), 1831–1839 (2012)

[45] Payre, W., Cestac, J., Delhomme, P.: Intention to use a fully automated car: Attitudes and a priori acceptability. Transportation research part F: traffic psychology and behaviour **27**, 252–263 (2014)

[46] Ross, P.E.: Robot, you can drive my car. IEEE Spectrum **51**(6), 60–90 (2014)

[47] Atkins, W.: Research on the impacts of connected and autonomous vehicles

(cavs) on traffic flow. Stage 2: Traffic Modelling and Analysis Technical Report (2016)

[48] Maurer, M., Gerdes, J.C., Lenz, B., Winner, H.: Autonomous Driving: Technical, Legal and Social Aspects. Springer, ??? (2016)

[49] Bimbraw, K.: Autonomous cars: Past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology. In: 2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO), vol. 1, pp. 191–198 (2015). IEEE

[50] Chen, L., Li, Y., Huang, C., Li, B., Xing, Y., Tian, D., Li, L., Hu, Z., Na, X., Li, Z., *et al.*: Milestones in autonomous driving and intelligent vehicles: Survey of surveys. IEEE Transactions on Intelligent Vehicles **8**(2), 1046–1056 (2022)

[51] Katrakazas, C., Quddus, M., Chen, W.-H., Deka, L.: Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. Transportation Research Part C: Emerging Technologies **60**, 416–442 (2015)

[52] Paden, B., Čáp, M., Yong, S.Z., Yershov, D., Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Transactions on intelligent vehicles **1**(1), 33–55 (2016)

[53] Ferguson, D., Howard, T.M., Likhachev, M.: Motion planning in urban environments. Journal of Field Robotics **25**(11-12), 939–960 (2008)

[54] Pivtoraiko, M., Kelly, A.: Differentially constrained motion replanning using state lattices with graduated fidelity. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2611–2616 (2008). IEEE

[55] Werling, M., Kammel, S., Ziegler, J., Gröll, L.: Optimal trajectories for time-critical street scenarios using discretized terminal manifolds. The International Journal of Robotics Research **31**(3), 346–359 (2012)

[56] LaValle, S.M., Kuffner Jr, J.J.: Randomized kinodynamic planning. The international journal of robotics research **20**(5), 378–400 (2001)

[57] Arslan, O., Berntorp, K., Tsiotras, P.: Sampling-based algorithms for optimal motion planning using closed-loop prediction. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 4991–4996 (2017). IEEE

[58] Falcone, P., Borrelli, F., Asgari, J., Tseng, H.E., Hrovat, D.: Predictive active steering control for autonomous vehicle systems. IEEE Transactions on control systems technology **15**(3), 566–580 (2007)

[59] Schwarting, W., Alonso-Mora, J., Pauli, L., Karaman, S., Rus, D.: Parallel autonomy in automated vehicles: Safe motion generation with minimal intervention. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 1928–1935 (2017). IEEE

[60] Liniger, A., Domahidi, A., Morari, M.: Optimization-based autonomous racing of 1: 43 scale rc cars. Optimal Control Applications and Methods **36**(5), 628–647 (2015)

[61] Varshney, K.R., Alemzadeh, H.: On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. Big data **5**(3), 246–255 (2017)

[62] Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011, pp. 1521–1528 (2011). IEEE

[63] Anderson, S.J., Karumanchi, S.B., Iagnemma, K., Walker, J.M.: The intelligent copilot: A constraint-based approach to shared-adaptive control of ground vehicles. IEEE Intelligent Transportation Systems Magazine **5**(2), 45–54 (2013)

[64] Buehler, M., Iagnemma, K., Singh, S.: The DARPA Urban Challenge: Autonomous Vehicles in City Traffic vol. 56. springer, ??? (2009)

[65] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., *et al.*: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2446–2454 (2020)

[66] Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C.R., Zhou, Y., *et al.*: Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9710–9719 (2021)

[67] Zhao, W., Queralta, J.P., Westerlund, T.: Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 737–744 (2020). IEEE

[68] Article 3.7 testing of autonomous vehicles (2022)

[69] Omeiza, D., Webb, H., Jirotka, M., Kunze, L.: Explanations in autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems **23**(8), 10142–10162 (2021)

[70] Cunneen, M., Mullins, M., Murphy, F.: Autonomous vehicles and embedded artificial intelligence: The challenges of framing machine driving decisions. Applied Artificial Intelligence **33**(8), 706–731 (2019)

[71] Ha, T., Kim, S., Seo, D., Lee, S.: Effects of explanation types and perceived risk on trust in autonomous vehicles. Transportation research part F: traffic psychology and behaviour **73**, 271–280 (2020)

[72] Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., Nass, C.: Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. International Journal on Interactive Design and Manufacturing (IJIDeM) **9**, 269–275 (2015)

[73] Board, N.T.S.: Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator: Mountain View, California, March 23, 2018. National Transportation Safety Board Washington, DC, USA (2020)

[74] Tabassi, E.: Artificial intelligence risk management framework (ai rmf 1.0) (2023)

[75] Hussain, R., Zeadally, S.: Autonomous cars: Research results, issues, and future challenges. IEEE Communications Surveys & Tutorials **21**(2), 1275–1313 (2018)

[76] Hoffman, R.R., Klein, G.: Explaining explanation, part 1: theoretical foundations. IEEE Intelligent Systems **32**(3), 68–73 (2017)

[77] Omeiza, D., Kollnig, K., Web, H., Jirotka, M., Kunze, L.: Why not explain? effects of explanations on human perceptions of autonomous driving. In: 2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO), pp. 194–199 (2021). IEEE

[78] Hergeth, S., Lorenz, L., Vilimek, R., Krems, J.F.: Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. Human factors **58**(3), 509–519 (2016)

[79] Payre, W., Cestac, J., Delhomme, P.: Fully automated driving: Impact of trust and practice on manual control recovery. Human factors **58**(2), 229–241 (2016)

[80] Rajaonah, B., Anceaux, F., Vienne, F.: Trust and the use of adaptive cruise control: a study of a cut-in situation. Cognition, Technology & Work **8**(2), 146–155 (2006)

[81] Muir, B.M., Moray, N.: Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. Ergonomics **39**(3), 429–460 (1996)

[82] Montavon, G., Samek, W., Müller, K.-R.: Methods for interpreting and understanding deep neural networks. Digital signal processing **73**, 1–15 (2018)

[83] Bastani, O., Kim, C., Bastani, H.: Interpreting blackbox models via model extraction. arXiv preprint arXiv:1705.08504 (2017)

[84] Yessenalina, A., Choi, Y., Cardie, C.: Automatically generating annotator rationales to improve sentiment classification. In: Proceedings of the ACL 2010 Conference Short Papers, pp. 336–341 (2010)

[85] Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, pp. 3–19 (2016). Springer

[86] Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing neural predictions. arXiv preprint arXiv:1606.04155 (2016)

[87] Kraus, J., Scholz, D., Stiegemeier, D., Baumann, M.: The more you know: trust dynamics and calibration in highly automated driving and the effects of takeovers, system malfunction, and system transparency. Human factors **62**(5), 718–736 (2020)

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

[89] M. Faas, S., Kraus, J., Schoenhals, A., Baumann, M.: Calibrating pedestrians' trust in automated vehicles: does an intent display in an external hmi support trust calibration and safe crossing behavior? In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–17 (2021)

[90] The global guide to autonomous vehicles (2022)

[91] Kalra, N., Paddock, S.M.: Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transportation Research Part A: Policy and Practice **94**, 182–193 (2016)

[92] Sadigh, D., Sastry, S., Seshia, S.A., Dragan, A.D.: Planning for autonomous cars that leverage effects on human actions. In: Robotics: Science and Systems, vol. 2, pp. 1–9 (2016). Ann Arbor, MI, USA

[93] Trautman, P., Ma, J., Murray, R.M., Krause, A.: Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation. The Int. Journal of Robotics Research **34**(3), 335–356 (2015)

[94] Kretzschmar, H., Spies, M., Sprunk, C., Burgard, W.: Socially compliant mobile robot navigation via inverse reinforcement learning. The International Journal of Robotics Research **35**(11), 1289–1307 (2016)

[95] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 961–971

(2016)

[96] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[97] Nayakanti, N., Al-Rfou, R., Zhou, A., Goel, K., Refaat, K.S., Sapp, B.: Wayformer: Motion forecasting via simple & efficient attention networks. arXiv preprint arXiv:2207.05844 (2022)

[98] Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)

[99] Strawn, K.J., Ayanian, N., Lindemann, L.: Conformal predictive safety filter for rl controllers in dynamic environments. arXiv preprint arXiv:2306.02551 (2023)

[100] Luo, R., Zhao, S., Kuck, J., Ivanovic, B., Savarese, S., Schmerling, E., Pavone, M.: Sample-efficient safety assurances using conformal prediction. In: Algorithmic Foundations of Robotics XV: Proceedings of the Fifteenth Workshop on the Algorithmic Foundations of Robotics, pp. 149–169 (2022). Springer

[101] Dietterich, T.G., Hostetler, J.: Conformal prediction intervals for markov decision process trajectories. arXiv preprint arXiv:2206.04860 (2022)

[102] Bortolussi, L., Cairoli, F., Paoletti, N., Smolka, S.A., Stoller, S.D.: Neural predictive monitoring. In: Runtime Verification: 19th Int. Conf., RV 2019, Porto, Portugal, October 8–11, 2019, Proceedings 19, pp. 129–147 (2019). Springer

[103] Fan, C., Qin, X., Xia, Y., Zutshi, A., Deshmukh, J.: Statistical verification of autonomous systems using surrogate models and conformal inference. arXiv preprint arXiv:2004.00279 (2020)

[104] Chen, Y., Rosolia, U., Fan, C., Ames, A., Murray, R.: Reactive motion planning with probabilisticsafety guarantees. In: Conf. on Robot Learning, pp. 1958–1970 (2021). PMLR

[105] Everett, M., Chen, Y.F., How, J.P.: Collision avoidance in pedestrian-rich environments with deep reinforcement learning. IEEE Access **9**, 10357–10377 (2021)

[106] Rawlings, J.B.: Tutorial overview of model predictive control. IEEE control systems magazine **20**(3), 38–52 (2000)

[107] Khatib, O.: Real-time obstacle avoidance for manipulators and mobile robots. The Int. journal of robotics research **5**(1), 90–98 (1986)

[108] Ferrer, G., Garrell, A., Sanfeliu, A.: Social-aware robot navigation in urban environments. In: 2013 European Conf. on Mobile Robots, pp. 331–336 (2013).

IEEE

[109] Van Den Berg, J., Guy, S.J., Lin, M., Manocha, D.: Reciprocal n-body collision avoidance. In: Robotics Research: The 14th Int. Symposium ISRR, pp. 3–19 (2011). Springer

[110] Chen, Y.F., Everett, M., Liu, M., How, J.P.: Socially aware motion planning with deep reinforcement learning. In: 2017 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 1343–1350 (2017). IEEE

[111] Chen, Y.F., Liu, M., Everett, M., How, J.P.: Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In: 2017 IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 285–292 (2017). IEEE

[112] Everett, M., Chen, Y.F., How, J.P.: Motion planning among dynamic, decision-making agents with deep reinforcement learning. In: 2018 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 3052–3059 (2018). IEEE

[113] Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)

[114] Kim, B., Pineau, J.: Socially adaptive path planning in human environments using inverse reinforcement learning. Int. Journal of Social Robotics **8**, 51–66 (2016)

[115] Pfeiffer, M., Schwesinger, U., Sommer, H., Galceran, E., Siegwart, R.: Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models. In: 2016 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 2096–2101 (2016). IEEE

[116] Tai, L., Paolo, G., Liu, M.: Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In: 2017 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 31–36 (2017). IEEE

[117] Farid, A., Veer, S., Ivanovic, B., Leung, K., Pavone, M.: Task-relevant failure detection for trajectory predictors in autonomous vehicles. In: Conf. on Robot Learning, pp. 1959–1969 (2023). PMLR

[118] Long, P., Fan, T., Liao, X., Liu, W., Zhang, H., Pan, J.: Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning. In: 2018 IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 6252–6259 (2018). IEEE

[119] Mészáros, A., Alonso-Mora, J., Kober, J.: Trajflow: Learning the distribution over trajectories. arXiv preprint arXiv:2304.05166 (2023)

[120] Mesbah, A., Wabersich, K.P., Schoellig, A.P., Zeilinger, M.N., Lucia, S., Badgwell, T.A., Paulson, J.A.: Fusion of machine learning and mpc under uncertainty: What advances are on the horizon? In: 2022 American Control Conf. (ACC), pp. 342–357 (2022). IEEE

[121] Wabersich, K.P., Hewing, L., Carron, A., Zeilinger, M.N.: Probabilistic model predictive safety certification for learning-based control. IEEE Transactions on Automatic Control **67**(1), 176–188 (2021)

[122] Lindemann, L., Cleaveland, M., Shim, G., Pappas, G.J.: Safe planning in dynamic environments using conformal prediction. IEEE Robotics and Automation Letters **8**(8), 5116–5123 (2023)

[123] Stankeviciute, K., M Alaa, A., Schaar, M.: Conformal time-series forecasting. Advances in Neural Information Processing Systems **34**, 6216–6228 (2021)

[124] Berkenkamp, F., Schoellig, A.P.: Safe and robust learning control with gaussian processes. In: 2015 European Control Conf. (ECC), pp. 2496–2501 (2015). IEEE

[125] Thrun, S.: Probabilistic robotics. Communications of the ACM **45**(3), 52–57 (2002)

[126] Niu, L., Zhang, H., Clark, A.: Safety-critical control synthesis for unknown sampled-data systems via control barrier functions. In: 2021 60th IEEE Conf. on Decision and Control (CDC), pp. 6806–6813 (2021). IEEE

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

[128] Rober, N., Katz, S.M., Sidrane, C., Yel, E., Everett, M., Kochenderfer, M.J., How, J.P.: Backward reachability analysis of neural feedback loops: Techniques for linear and nonlinear systems. IEEE Open Journal of Control Systems (2023)

[129] Muntwiler, S., Wabersich, K.P., Carron, A., Zeilinger, M.N.: Distributed model predictive safety certification for learning-based control. IFAC-PapersOnLine **53**(2), 5258–5265 (2020)

[130] Papadopoulos, H.: Inductive conformal prediction: Theory and application to neural networks. In: Tools in Artificial Intelligence. Citeseer, ??? (2008)

[131] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, ??? (2005)

[132] Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning Research **9**(3) (2008)

[133] Fontana, M., Zeni, G., Vantini, S.: Conformal prediction: a unified review of

theory and new challenges. Bernoulli **29**(1), 1–23 (2023)

[134] Foffano, D., Russo, A., Proutiere, A.: Conformal off-policy evaluation in markov decision processes. arXiv preprint arXiv:2304.02574 (2023)

[135] Taufiq, M.F., Ton, J.-F., Cornish, R., Teh, Y.W., Doucet, A.: Conformal off-policy prediction in contextual bandits. arXiv preprint arXiv:2206.04405 (2022)

[136] Dixit, A., Lindemann, L., Wei, S., Cleaveland, M., Pappas, G.J., Burdick, J.W.: Adaptive conformal prediction for motion planning among dynamic agents. arXiv preprint arXiv:2212.00278 (2022)

[137] Muthali, A., Shen, H., Deglurkar, S., Lim, M.H., Roelofs, R., Faust, A., Tomlin, C.: Multi-agent reachability calibration with conformal prediction. arXiv preprint arXiv:2304.00432 (2023)

[138] Garcıa, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research **16**(1), 1437–1480 (2015)

[139] Berkenkamp, F., Turchetta, M., Schoellig, A., Krause, A.: Safe model-based reinforcement learning with stability guarantees. Advances in neural information processing systems **30** (2017)

[140] Koller, T., Berkenkamp, F., Turchetta, M., Krause, A.: Learning-based model predictive control for safe exploration. In: 2018 IEEE Conf. on Decision and Control (CDC), pp. 6059–6066 (2018). IEEE

[141] Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., Tassa, Y.: Safe exploration in continuous action spaces. arXiv preprint arXiv:1801.08757 (2018)

[142] Glossop, C.R., Panerati, J., Krishnan, A., Yuan, Z., Schoellig, A.P.: Characterising the robustness of reinforcement learning for continuous control using disturbance injection. arXiv preprint arXiv:2210.15199 (2022)

[143] Tearle, B., Wabersich, K.P., Carron, A., Zeilinger, M.N.: A predictive safety filter for learning-based racing control. IEEE Robotics and Automation Letters **6**(4), 7635–7642 (2021)

[144] Wabersich, K.P., Zeilinger, M.N.: A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. Automatica **129**, 109597 (2021)

[145] Seto, D., Krogh, B., Sha, L., Chutinan, A.: The simplex architecture for safe online control system upgrades. In: Proceedings of the 1998 American Control Conf.. ACC (IEEE Cat. No. 98CH36207), vol. 6, pp. 3504–3508 (1998). IEEE

[146] Prajna, S., Jadbabaie, A.: Safety verification of hybrid systems using barrier certificates. In: HSCC, vol. 2993, pp. 477–492 (2004). Springer

[147] Wieland, P., Allgöwer, F.: Constructive safety using control barrier functions. IFAC Proceedings Volumes **40**(12), 462–467 (2007)

[148] Wabersich, K.P., Zeilinger, M.N.: Predictive control barrier functions: Enhanced safety mechanisms for learning-based control. IEEE Transactions on Automatic Control (2022)

[149] Taylor, A., Singletary, A., Yue, Y., Ames, A.: Learning for safety-critical control with control barrier functions. In: Learning for Dynamics and Control, pp. 708–717 (2020). PMLR

[150] Didier, A., Jacobs, R.C., Sieber, J., Wabersich, K.P., Zeilinger, M.N.: Approximate predictive control barrier functions using neural networks: A computationally cheap and permissive safety filter. arXiv preprint arXiv:2211.15104 (2022)

[151] Didier, A., Wabersich, K.P., Zeilinger, M.N.: Adaptive model predictive safety certification for learning-based control. In: 2021 60th IEEE Conf. on Decision and Control (CDC), pp. 809–815 (2021). IEEE

[152] Leeman, A.P., Köhler, J., Benanni, S., Zeilinger, M.N.: Predictive safety filter using system level synthesis. arXiv preprint arXiv:2212.02111 (2022)

[153] Cauchois, M., Gupta, S., Ali, A., Duchi, J.C.: Robust validation: Confident predictions even when distributions shift. arXiv preprint arXiv:2008.04267 (2020)

[154] Zaffran, M., Féron, O., Goude, Y., Josse, J., Dieuleveut, A.: Adaptive conformal predictions for time series. In: International Conference on Machine Learning, pp. 25834–25866 (2022). PMLR

[155] Lasota, P.A., Fong, T., Shah, J.A., *et al.*: A survey of methods for safe human-robot interaction. Foundations and Trends® in Robotics **5**(4), 261–349 (2017)