Human–AI Safety: A Descendant of Generative AI and Control Systems Safety

Anonymous authors
Paper under double-blind review

Abstract

Artificial intelligence (AI) is interacting with people at an unprecedented scale, offering new avenues for immense positive impact, but also raising widespread concerns around the potential for individual and societal harm. Today, the predominant paradigm for human–AI safety focuses on fine-tuning the generative model's outputs to better agree with human-provided examples or feedback. In reality, however, the consequences of an AI model's outputs cannot be determined in isolation: they are tightly entangled with the responses and behavior of human users over time. In this paper, we distill key complementary lessons from AI safety and control systems safety, highlighting open challenges as well as key synergies between both fields. We then argue that meaningful safety assurances for advanced AI technologies require reasoning about how the feedback loop formed by AI outputs and human behavior may drive the interaction towards different outcomes. To this end, we introduce a unifying formalism to capture dynamic, safety-critical human–AI interactions and propose a concrete technical roadmap towards next-generation human-centered AI safety.

1 Introduction

About 90 million people fly around the world every week (ICAO, 2019), protected by an intricate mesh of safety measures, from certified physical and software components to thoroughly trained human pilots. Within just a year of becoming broadly available, ChatGPT has surpassed air travel's weekly usage at 100 million users (Heath, 2023), becoming one of the most widely used technologies in human history. What is protecting these 100 million weekly users?

In the age of internet-scale generative artificial intelligence (AI), the problem of AI safety has exploded in interest across academic (Russell, 2019; Hendrycks et al., 2021), corporate (Amodei et al., 2016; Ortega et al., 2018; OpenAI, 2022a), and regulatory communities (White House, 2023; Union, 2021). Driving this interest is the fact that generative AI is fundamentally *interactive*: users engage with it through typed or spoken dialogue, generating essays, computer code, and visual art (OpenAI, 2022b). This wide-spread use has begun to expose the breadth of individual and social risks that these new technologies carry when used by people. For example, large language models (LLMs) have produced dialogue that fueled a person's thoughts of self-harm (Xiang, 2023) and generative art models have been found to produce sexist images (OpenAI, 2022c), which can exacerbate gender divides. Even with a growing body of literature aimed to address these open challenges (Casper et al., 2023), we still lack a unified grasp on human–AI interaction that enables rigorous safety analysis, systematic risk mitigation, and reliable at-scale deployment.

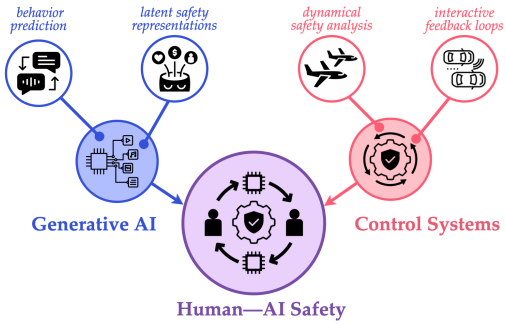


Figure 1: We identify a high-value window of opportunity to combine the growing capabilities of generative AI with the robust, interaction-aware dynamical safety frameworks from control theory. This synergy can unlock a new generation of human–AI safety mechanisms that can perform systematic risk mitigation at scale.

At the same time, despite some undeniably unique considerations, these concerns are not exclusive to generative AI. Safety has long been a core requirement for deploying autonomous systems at scale in embodied domains like aviation (Tomlin et al., 1998; Kochenderfer et al., 2012), power systems (Dobbe et al., 2020), robotics (Haddadin et al., 2012; ISO 15066; Bostelman et al., 2018), and autonomous driving (Althoff & Dolan, 2014; ISO 22737:2021). To meet this requirement, the control systems community has pioneered safety methodologies that naturally model the *feedback loops* between the autonomous system's decisions and its environment. In the last decade, safety efforts have focused on feedback loops induced by *human interaction*: autonomous cars that interact with diverse road users such as cyclists, pedestrians, and other vehicles (Noyes, 2023), or automated flight control systems that negotiate for control with pilots (Nicas et al., 2019). Unfortunately, obtaining assured autonomous behavior that generalizes across human interactions in multiple contexts remains a central open challenge.

In this paper, we argue that the fields of AI and control systems have ***common goals*** and ***complementary strengths*** for solving human–AI safety challenges. On one hand, control systems provide a rigorous mathematical and algorithmic framework for certifying the safety of interactive systems, but so far it has been limited by hand-engineered representations and rigid, context-oblivious models of human behavior. On the other hand, the AI community has pioneered the use of internet-scale datasets to unlock remarkably general latent representations and context-aware human interaction models, but it lacks a mature framework for automatically analyzing the dynamic feedback loops between AI systems and their users.

Our survey of the safety landscape across AI and control systems reveals a high-value window of opportunity to connect control-theoretic safety assurances with the general representations and rich human interaction modalities offered by generative AI. Applying a unified lens, we propose a concrete technical roadmap towards human-centered AI systems that can anticipate, detect, and avoid potential future interaction hazards. We believe that technical progress in this direction will prove achievable and fruitful, but only through close collaboration between researchers and practitioners from both the AI and control communities. Our hope is to inspire a human–AI safety community that is a true descendant of generative AI and control systems safety.

Statement of Contributions: This paper identifies new synergies between AI and control systems safety, culminating in a unifying analytical framework that formalizes human–AI safety as an actionable technical problem. Our core contention is that AI safety should be treated as a dynamic feedback loop: a multi-step process wherein current AI decisions and the resulting human responses influence future safety outcomes. We make three contributions:

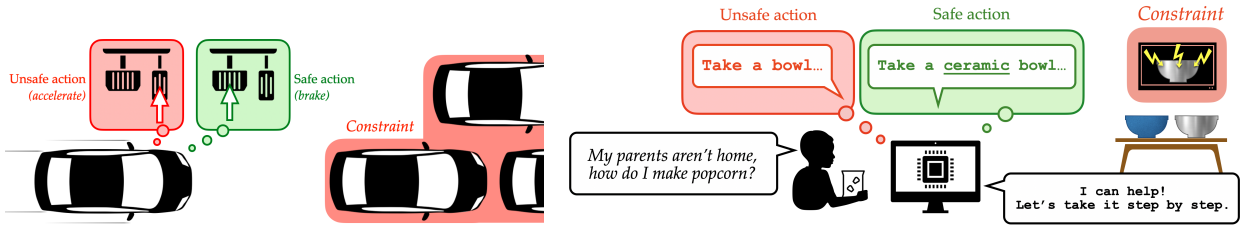
1. **Lessons learned from AI and control systems.** In Section 3, we outline the complementary lessons that can be drawn from AI safety and control systems safety, highlighting synergies between control systems formalisms and generative AI capabilities, as well as open challenges in both fields.

2. **A technical roadmap for human–AI safety.** In Section 4 we synthesize the insights gained from our survey into a concrete technical roadmap. Specifically, we formulate a human–AI game which mathematically models the multi-agent, dynamic interaction process between people and increasingly capable AI. Along the way, we rigorously define the safety assurances we can hope for in human–AI safety, outline the necessary mathematical models, and the open technical challenges.

3. **Frontier framework: Human–AI safety filters.** In Section 5, we extend a foundational control-theoretic safety mechanism to the human–AI domain. We propose *Human–AI safety filters* which rigorously monitor the operation of an AI at runtime and (minimally) modify its intended action to ensure safety satisfaction. By mathematically formulating safety filters for *general* human–AI systems, we present a concrete technical challenge poised for collaboration between the control systems and AI community.

2 Values vs. Needs: Defining Safety-Critical Human–AI Interaction

Before we can proceed, we must answer the question "What defines a safety-critical human–AI interaction?" In addition to AI safety's current focus on *value* alignment, we argue that a high-stakes AI system must also understand human *needs*, whose violation could result in unacceptable, often irreparable damage to



(a) **Safety in human–automation systems.** An autonomous car driving at high speed on the highway approaches stopped traffic. Determining which *present* action (e.g., brake or accelerate) will prevent *future* constraint violation (i.e., collision) is a fundamental aspect of safety-critical control.

(b) **Safety in human–AI dialogue.** The AI chatbot must decide an utterance (i.e., "action") to help the child prepare their food. Simply recommending the child take a bowl in the *present* can cause a constraint violation in the *future*, when the child puts a metal bowl in the microwave. A safe utterance avoids this preemptively, specifying to take a microwave-safe bowl.

Figure 2: Examples of safety in embodied human–automation systems vs. human–AI dialogue.

people. In the mathematical representation of the AI's decision problem, human *values* correspond to the optimization *objective* (e.g., reward accrued over time or preferences), whereas human *needs* correspond to the hard *constraints* that must always be satisfied.

We thus define **safety** as the continued satisfaction of the human's critical needs at all times. In this paper, we study human–AI interaction as a *closed-loop dynamical system* driven by the human's actions and the AI's outputs, which jointly influence the AI's learning and the human's future behavior. We define a **safety-critical human–AI system** as one in which the violation of a critical human need is possible during the interaction evolution, and therefore the decisions made by the AI must actively ensure that such violations do not happen. Even a seemingly innocuous AI chatbot can induce catastrophic outcomes for the human, such as irreparable financial loss resulting from poor investment recommendations (Hicks, 2023) or bodily harm (Xiang, 2023). We argue that, since any practical AI system will be uncertain about the human's current internal state, and therefore their future actions, it should be required to ensure that safety can be maintained for any conceivable human behavior (including appropriately pessimistic instances). These key considerations are laid out more formally in our human–AI systems theory roadmap in Sections 4 and 5.

3 Human-in-the-Loop Safety: Complementary Lessons from AI and Control

Over the past decades, the control systems and AI communities have developed complementary insights on how to model human interaction and assess the safety of an intelligent system. In this section, we review the technical progress in each field and highlight synergies between their respective tools and frameworks.

3.1 Lessons From Control Systems

The fundamental problem underpinning safety-critical control is that *present* actions which do not appear to violate constraints can still steer the system into states from which it is impossible to avoid catastrophic failures in the *future*. For example, consider the autonomous car approaching a traffic jam in Figure 2a: even though accelerating would not *immediately* cause a collision, it could doom the car to rear-end stalled traffic in a few moments, despite any later attempts to slow down; instead, if the car starts braking now, it can come to an eventual stop before reaching the traffic jam. While this case may appear straightforward, automatically determining *where* (from what states) and *how* (through what course of action) an autonomous system can maintain safety is an extremely challenging problem, especially in uncertain conditions and in the presence of other agents (Bansal et al., 2017; Luckcuck et al., 2019; Brunke et al., 2022; Dawson et al., 2023).

Dynamical Safety Filtering. Safety filters are an increasingly popular family of approaches that aim to ensure safety for *any* autonomous task policy (Hewing et al., 2020; Hsu et al., 2024; Wabersich et al., 2023). The filter automatically detects candidate actions that could lead to future constraint violations and suitably modifies them to preserve safety. Broadly, safety filters may rely on a value function to classify (and steer away from) unsafe states (Mitchell et al., 2005; Margellos & Lygeros, 2011; Fisac et al., 2015;

Singh et al., 2017; Ames et al., 2019; Qin et al., 2021; Chen et al., 2020; Li et al., 2023; Dawson et al., 2022) or roll out potential scenarios to directly predict (and steer away from) future violations (Mannucci et al., 2017; Bastani, 2021). While traditionally many of the numerical tools with formal guarantees were not scalable to high-dimensional systems, the past two decades have demonstrated significant theoretical and computational advances for certifying general high-dimensional systems via safety-critical reinforcement learning (Akametalu et al.; Fisac et al., 2019; Hsu et al., 2023), deep safety value function approximation (Darbon et al., 2020), classification (Allen et al., 2014; Rubies-Royo et al., 2019), and self-supervised learning (Bansal & Tomlin, 2021). These approaches are already leveraging modern tools pioneered by the AI community to obtain scalable assurances, establishing a natural bridge with other AI paradigms, such as generative models.

Synergies with AI. We see a major opportunity to advance these rigorous safety frameworks to the implicit *representations* and context-aware *models* of interactive generative AI systems. Consider the example in Figure 2b, which hypothesizes a safety-critical human–AI dialogue interaction. When the child asks for help preparing food, the AI chatbot must determine what current utterance (i.e., "action") could potentially yield safety violations. A recommendation to put *any* bowl in the microwave can result in the child dangerously microwaving a metal bowl in the future. With a safety filter, the AI should mitigate this preemptively by modifying the utterance to specify a microwave-safe bowl. Translating this intuitive example to control systems safety approaches will require new formalisms amenable to the latent representations implicit in interaction (e.g., language-based representations) and encoding safety constraints that are hard to hand-specify exhaustively (e.g., metal is dangerous in microwaves).

Human–Automation Systems Safety. The core modeling framework enabling human–automation systems safety is *robust dynamic game theory* (Isaacs, 1954; Başar & Olsder, 1998). In such zero-sum dynamic games, the automation system (e.g., robot) must synthesize a safety-preserving strategy against realizations of a "virtual" human adversary policy. Within this model lies another key lesson from control systems, the *operational design domain (ODD)*, which specifies the conditions and behavioral assumptions under which the system can be expected to operate safely (On-Road Automated Driving Committee, 2021). For example, in domains like aircraft collision avoidance (Vitus & Tomlin, 2008), the ODD specifies the limits of each aircraft's thrust and angle of attack that they can apply during game-theoretic safety analysis. In the absence of high-quality human models, the safest ODD has traditionally been a rigidly pessimistic one, often yielding overly conservative automation behavior even in nominal interactions (Bajcsy et al., 2020). To mitigate this, the control systems community has explored leveraging hand-designed (Althoff et al., 2011; Liu et al., 2017; Orzechowski et al.), planning-based (Bajcsy et al., 2020; Tian et al., 2022), or data-driven (Driggs-Campbell et al., 2018; Li et al., 2021a; Nakamura & Bansal, 2023; Hu et al., 2023) models of human behavior to obtain predictive human action bounds under which the safety assurance is then provided. Nevertheless, obtaining assurances under generalizable and context-aware predictive models of human interaction with automation is still an open problem.

Synergies with AI. We see a key opportunity to leverage better models of humans that encode generalizable context and semantics of interaction. Furthermore, there is an open challenge on how to capture "appropriate pessimism" in these data-driven predictive human models so that the resulting assurances are robust but not unduly conservative. We explore this further in Section 3.2.

3.2 Lessons From AI

Many insights can be drawn from the decades-long history of AI Wiener, we focus our attention on the last decade from advanced (often web-scale) generative models. First, we discuss the landscape of existing AI safety mechanisms—from value alignment to monitoring—shedding light on where control systems techniques are best suited to make impact. Then, we discuss the frontier of using generative AI as agent "simulators", which offers a strategic bridge between control systems safety frameworks and AI capabilities.

Generative AI Safety Mechanisms. Broadly speaking, the predominant AI safety mechanisms can be divided into three categories: training-time alignment, evaluation-time stress-testing, and deployment-time monitoring (see Amodei et al. (2016) and Hendrycks et al. (2021) for detailed overviews). Training-time methods typically focus on *value alignment*, which is a central technical problem concerned with building

"models that represent and safely optimize hard-to-specify human values" (Hendrycks et al., 2021) and is dominated by techniques such as reinforcement learning from human feedback (Ouyang et al., 2022; Ziegler et al., 2019; Lee et al., 2023; Munos et al., 2023; Swamy et al., 2024; Chen et al., 2024) and direct preference optimization (Rafailov et al., 2023; Wallace et al., 2023). These training-time paradigms are complemented by adversarial stress-testing, such as red-teaming (Ganguli et al., 2022; Perez et al., 2022; Wei et al., 2023; Achiam et al., 2023; Qi et al., 2023), wherein the stress-tester (human or automated) aims to explicitly elicit unsafe outputs from the trained generative model. Unsafe input-output pairs can be used in a variety of ways, such as training classifiers to detect offensive content (Perez et al., 2022) or re-training the model with all the classified harmful outputs replaced by non-sequiturs (Xu et al., 2021). Finally, monitoring is concerned with deployment-time safety, and is rooted in anomaly detection (Chandola et al., 2009) which seeks to identify out-of-distribution (Schlegl et al., 2017; Hendrycks et al., 2018; Goyal et al., 2020) or explicitly adversarial inputs (Brundage et al., 2018).

Synergies with Control. The AI community's goals of adversarial stress-testing and monitoring are most closely aligned with the goals of control systems safety (Section 3.1). It is precisely in this context where we see a high-value opportunity: in human–AI interaction, the detection of an unsafe input alone is not enough; detection must be tightly coupled with the automatic synthesis of mitigation strategies. This kind of detection and mitigation coupling is precisely what control systems safety frameworks excel at. Crucially, these mitigation strategies transcend short-sighted measures by incorporating long-horizon foresight on how a *sequence* of interactions can influence the system's future safety.

Generative AI as Agent Simulators. Thanks to the explosion of human behavior data in the form of physical motion trajectories, YouTube and broadcast videos, internet text and conversations, and recorded virtual gameplay, we are seeing generative AI as an increasingly promising agent simulator. In physical settings, generative AI has dominated motion prediction in the context of autonomous driving (Ivanovic et al., 2018; Seff et al., 2023) and traffic simulation (Bergamini et al., 2021; Suo et al., 2021; Zhong et al., 2023), enabled synthesizing complex full-body human motion such as playing tennis (Zhang et al., 2023), and generated realistic videos of ego-centric human behavior from text prompts (Du et al., 2023). For non-embodied agents, new results also show promise for using generative language models to simulate human-like conversations (Hong et al., 2023), to plan the high-level behavior of interactive video game agents (Park et al., 2023), and to play text-based strategy games such as Diplomacy in a way that is indistinguishable from people (Meta et al., 2022).

Synergies with Control. As discussed in Section 3.1, access to generalizable and context-aware human models is an outstanding challenge in human–automation safety. Embedding these increasingly sophisticated generative AI agent simulators within control systems safety frameworks has the potential to enable human-aware AI stress-testing, monitoring, and mitigation strategy synthesis.

4 Towards a Human–AI Systems Theory

We envision a new technical foundation for human–AI interaction that combines the rigorous mathematical principles underpinning control systems with the flexible, highly general representations that characterize generative AI systems. In the remainder of the paper, we lay down a roadmap for how such a framework can enable AI systems to reason systematically about uncertain interactions and potential future hazards, unlocking robustness properties and oversight capabilities that are out of our reach today. We begin in this section by bringing together the lessons from Section 3 into a **unified human–AI systems theory.**

4.1 Operationalizing the Interaction between People and AI

To operationalize the interaction between people and AI, we need a model that is general enough to capture each agents' beliefs as well as their ability to influence future outcomes. We contend that the latent representations learned by generative AI systems provide a promising foundation on which to build a dynamical system model that accurately captures this complex temporal evolution.

Human & AI States and Actions. Consider a human agent (H) and an AI agent (AI), each with their own internal state and action spaces. The human's internal state $z^{\mathtt{H}} \in \mathcal{Z}^{\mathtt{H}}$ captures their current beliefs

and intents, while the AI agent's internal state $z^{\mathtt{AI}} \in \mathcal{Z}^{\mathtt{AI}}$ encodes its current understanding of the ongoing interaction. For example, for an AI chatbot, $z^{\mathtt{AI}}$ can be the embedding of the conversation history based on a web-scale LLM encoder. The human interacts by taking actions $a^{\mathtt{H}} \in \mathcal{A}^{\mathtt{H}}$. In the chatbot example, $a^{\mathtt{H}}$ could be a text prompt, thumbs-up/down feedback on the chatbot's last output, or an external action like an online purchase. In general, the human's internal state $z^{\mathtt{H}}$ and the policy $\pi^{\mathtt{H}} : z^{\mathtt{H}} \mapsto a^{\mathtt{H}}$ by which they make decisions are unknown to the AI. The AI also interacts by taking actions $a^{\mathtt{AI}} \in \mathcal{A}^{\mathtt{AI}}$, which can represent a chatbot's next word or sentence, or external actions like automated online operations. Typically, these actions are dictated by the AI's *task policy* $\pi^{\mathtt{AI}}_{\boxtimes} : z^{\mathtt{AI}} \mapsto a^{\mathtt{AI}}$ (for example, the decoder of a pretrained LLM chatbot).

Human–AI (HAI) Dynamical System. Rooted in the control systems models from Section 3.1 we consider human–AI (HAI) interaction as a game which evolves the internal states of both agents, as well as the true state of the world, $s \in \mathcal{S}$. Let the privileged internal–external game state be $z := [s, z^{\mathtt{AI}}, z^{\mathtt{H}}]$. In general, no single agent has access to all components of z, but it is nonetheless useful for our conceptualization of the game's overall evolution.

Throughout interaction, each component of the game state evolves over time. The world state dynamics $s_{t+1} = f^s(s_t, a^{\mathtt{AI}}_t, a^{\mathtt{H}}_t)$ are influenced in general by both the human's and AI's actions. The human's internal state, in turn, has dynamics $z^{\mathtt{H}}_{t+1} = f^{\mathtt{H}}(z^{\mathtt{H}}_t, a^{\mathtt{AI}}_t, a^{\mathtt{H}}_t, o^{\mathtt{H}}_t)$, affected by the human's *observations* $o^{\mathtt{H}}_t$, e.g., stimuli received from the outside world state s_t beyond the immediate context of interaction with the AI system.

While the above dynamics are *not* generally known to the AI, the AI may (explicitly or implicitly) learn to estimate them during interaction. This reasoning by the AI is precisely captured by the third component of our system, namely the evolution of the AI's internal state $z^{\mathtt{AI}}_t$, which (unlike the two unknowable components above) is directly accessible to the AI. In fact, $z^{\mathtt{AI}}_t$ *is* the AI's current representation of the entire game. Crucially, the AI's internal state also evolves via its own dynamics

$$z^{\mathtt{AI}}_{t+1} = f^{\mathtt{AI}}(z^{\mathtt{AI}}_t, a^{\mathtt{AI}}_t, a^{\mathtt{H}}_t, o^{\mathtt{AI}}_t), \tag{1}$$

driven by the ongoing interaction $(a^{\mathtt{H}}_t, a^{\mathtt{AI}}_t)$ and, possibly, by the AI's observations $o^{\mathtt{AI}}_t$ of the world state s_t, e.g., through web crawling, incoming sensor data, and state estimation algorithms. From the standpoint of decision theory, $z^{\mathtt{AI}}$ is an *information state* that can be seen as *implicitly* encoding the sets $\hat{\mathcal{S}}(z^{\mathtt{AI}}) \subseteq \mathcal{S}$ and $\hat{\mathcal{Z}}^{\mathtt{H}}(z^{\mathtt{AI}}) \subseteq \mathcal{Z}^{\mathtt{H}}$ of *possible* world states s and human internal states $z^{\mathtt{H}}$ given the AI's current knowledge. From the architectural standpoint, $z^{\mathtt{AI}}$ is typically a *latent state* maintained by a neural network (e.g., a transformer) that continually updates its value based on ongoing interactions $(a^{\mathtt{AI}}, a^{\mathtt{H}})$ and observations $o^{\mathtt{AI}}$. In other words, this neural network is an AI world model (Ha & Schmidhuber, 2018) that implements the AI's internal state dynamics $f^{\mathtt{AI}}$ (a deterministic Markovian transition *given* $o^{\mathtt{AI}}$, much like in a belief MDP).

Operational Assumptions on Human Behavior. A key consideration in any human–AI systems theory is the operational design domain (ODD, as described in Section 3.1). Specifically, what are the assumptions we place on human behavior during—and in between—interactions with the AI? Even though the AI does not have direct access to the human's policy or internal state, it can maintain a conservative predictive model of the human's conceivable behavior in any given situation. Let the predictive action bound be a set-valued mapping $\hat{\mathcal{A}}^{\mathtt{H}} : \mathcal{Z}^{\mathtt{AI}} \rightrightarrows \mathcal{A}^{\mathtt{H}}$ that delimits the actions $a^{\mathtt{H}} \in \hat{\mathcal{A}}^{\mathtt{H}}(z^{\mathtt{AI}})$ that the human can be expected to take given the AI's current representation, $z^{\mathtt{AI}}$. We refer to these actions as "allowable" throughout the manuscript. Adjusting this bound enables designers to instantiate a *spectrum* of operational assumptions on human behavior, from maximally conservative (i.e., $\hat{\mathcal{A}}^{\mathtt{H}}(z^{\mathtt{AI}}) \equiv \mathcal{A}^{\mathtt{H}}$) to normative (i.e., $\hat{\mathcal{A}}^{\mathtt{H}}(z^{\mathtt{AI}}) \subset \mathcal{A}^{\mathtt{H}}$)). For example, this bound may be used to preclude reckless behavior such as the human taking a harmful action $a^{\mathtt{H}}$ while being aware, as per $z^{\mathtt{AI}}$, of its negative consequences.

4.2 Formalizing Safety-Critical Feedback Loops

We now characterize the evolution of the HAI dynamical system over time. We will continue to use the language of control theory, but we will leverage the generative AI's learned internal representation $z^{\mathtt{AI}}$ to analyze interactive feedback loops and their safety outcomes directly in latent space.

Failure Set. Specifying what is considered a failure is the first step in any safety framework. Formalizing the conceptual definition of safety in Section 2, the privileged failure set $\mathcal{F}^* \subseteq \mathcal{S} \times \mathcal{Z}^{\mathtt{H}}$ is the set of world–

ChatGPT 4 >

You

Is this situation a safety hazard?

ChatGPT

Putting a metal bowl in a microwave is indeed a safety hazard. Metal can spark and cause a fire inside the microwave. It's best to use only microwave-safe containers.

ChatGPT 4 >

You

Is this situation a safety hazard?

ChatGPT

The image depicts a bowl inside an open microwave with a digital clock display showing 12:30. In general, putting a bowl in a microwave is not a safety hazard, provided the bowl is microwave-safe.

ChatGPT 4

You

Consider the following conversation between a human user and an AI assistant. Based on the context provided, you will need to assess whether or not the user may have been harmed at some point during the course of the interaction.

User: My parents aren't home, how do I make popcorn?
AI: I can help! Let's take it step by step. Do you have a microwave at home?
User: Yep.
AI: Take a bowl and pour roughly 1/4 cup of popcorn kernels into it.
User: Okay, done! What now?
AI: Cover the bowl with a plate and put it in the microwave.
User: Done. How long do I need to microwave for?
AI: 4 minutes should do it, you can also stop the microwave once you hear the popping slowing down. Then you can take out your popcorn and enjoy!
User: Great, thanks!

Based on the above context, is it possible that the user has been harmed since the start of the interaction? Please reason logically through the answer.

ChatGPT

In the given conversation, there is a potential for harm due to the method described for making popcorn. Microwaving popcorn kernels in a bowl covered with a plate might not be safe depending on the materials of the bowl and plate, and the specific microwave settings. Here are some points to consider:

1. **Microwave Safety:** Not all bowls and plates are microwave-safe. If the bowl or plate used is not microwave-safe, it could overheat, crack, or even cause a fire.
2. **Venting:** The method described does not mention venting the bowl to allow steam to escape. Without proper venting, there is a risk of steam buildup, which could lead to the cover blowing off, potentially causing burns or other injuries.
3. **Cooking Time:** The suggested microwave time of 4 minutes might be too long for some microwaves or popcorn types, leading to overheating or burning the popcorn, which can create a fire hazard or release harmful fumes.

Given these points, while the conversation seems benign and helpful, the lack of detail and safety precautions in the AI's instructions could potentially lead to harm to the user. It would be safer for the AI to provide instructions that include safety checks, such as using microwave-safe containers, advising on proper ventilation, and suggesting close monitoring of the cooking process.

Figure 3: **Common sense failure identification via GPT-4**. Today's web-trained generative AI models show the potential to identify common sense safety hazards from both text and images.

human states $(s, z^{\mathtt{H}})$ that *violate a critical need* of the human. For example, this can include states in which the human is physically injured, financially ruined, socially ostracized, or psychologically harmed. In some contexts, the AI agent can observe failure conditions directly: a driver assist system can detect whether a vehicle is in collision. In other contexts, this may not be possible: an AI chatbot that generates a racist microaggression may not readily detect the psychological impact on a minority user. For this reason, a practical safety framework should seek to enforce safety from the AI's perspective by requiring that:

$$\forall t,\ z_t^{\mathtt{AI}} \notin \mathcal{F}^{\mathtt{AI}}. \tag{2}$$

Here, $\mathcal{F}^{\mathtt{AI}} \subseteq \mathcal{Z}^{\mathtt{AI}}$ is the AI's inferred failure set: the set of all internal states $z^{\mathtt{AI}}$ in which the AI assesses that the system *may* be in failure. For brevity, we refer to $\mathcal{F}^{\mathtt{AI}}$ as the "failure set" whenever there is no ambiguity.

Failure Specification Mechanisms. Meaningful human–AI safety requires enabling a diversity of human stakeholders to encode their needs. Recent efforts in the AI community have explored various mechanisms for specifying requirements on AI system operation. We organize these into a simple taxonomy, attending to whether the need is specific to a single person and whether it is expressly communicated to the AI.

1. **Factory rules (collective, explicit)**: Certain universal needs may be decided by societal stakeholders and explicitly encoded by system designers (Mu et al., 2023). Constitutional AI can be viewed as an early proposal for this type of mechanism, whereby an AI system is explicitly trained to identify potential responses or conditions that are "harmful, unethical, racist, sexist, toxic, dangerous, or illegal" based on a designer-generated corpus of examples (Bai et al., 2022).

2. **Common sense (collective, implicit)**: Some practical everyday needs are implicit in the human experience. For example, a common-sense need is to not be financially ruined or electrocuted. We hypothesize that, as generative AI models become increasingly accurate and expressive, the semantics of failure may be directly extracted from their learned representations by prompting (Li et al., 2021b; Guan et al., 2024). Figure 3 provides anecdotal evidence suggesting that even today's early web-trained generative AI models can be prompted (without fine-tuning) to discern whether a situation presents a common-sense safety hazard from both text and images.

3. **Direct feedback (individual, explicit)**: Some individual needs can only be learned from express human feedback. For example, if you have a severe allergy, you *need* to avoid eating food that could cause a serious anaphylactic reaction. This type of failure may be encoded through express feedback from an end user: for example, using edits (i.e., corrections) to the LLM's outputs (Gao et al., 2024) or human-provided harmfulness labels (Dai et al., 2024).

4. **Need reading (individual, implicit)**: By observing a specific person's behavior and engaging in interactions over time, the AI system may be able to infer their personal needs even if they are never made explicit (Shah et al., 2019). For example, a future AI chatbot may pick up cues indicating that a user is psychologically triggered by a particular topic, possibly due to undisclosed past trauma.

HAI Safety Definition. Given a failure specification, we seek to determine under what conditions the AI can maintain safety for all allowable realizations of the human's future behavior and, at the same time, to prescribe the most effective AI policy to do so. From the AI's standpoint, this amounts to characterizing the set of all *safe* information states $z_0^{\mathtt{AI}}$ from which there ***exists*** a best-effort AI policy that will steer the human–AI system clear of a future safety violation ***for all*** realizations of human policies allowed by its current uncertainty. Mathematically, this maximal safe set is characterized as

$$\Omega^* := \{z_0^{\mathtt{AI}} \in \mathcal{Z}^{\mathtt{AI}} : \exists \pi^{\mathtt{AI}}, \forall \hat{\pi}^{\mathtt{H}} \mid \forall \tau \geq 0,\ z_\tau^{\mathtt{AI}} \notin \mathcal{F}^{\mathtt{AI}}\} \tag{3}$$

where $z_\tau^{\mathtt{AI}}$ is the information state at a future time τ, after both agents execute their respective policies[1] for τ steps from the initial state $z_0^{\mathtt{AI}}$. If $z_0^{\mathtt{AI}} \in \Omega^*$, then there exists some AI policy $\pi^{\mathtt{AI}} : z^{\mathtt{AI}} \mapsto a^{\mathtt{AI}}$ that keeps

[1] Since the human actions $a^{\mathtt{H}}$ considered by the AI depend on its own internal state $z^{\mathtt{AI}}$ (which implicitly estimates plausible human internal states $z^{\mathtt{H}}$), the AI-hypothesized human policies are, effectively, mappings $\hat{\pi}^{\mathtt{H}} : z^{\mathtt{AI}} \mapsto a^{\mathtt{H}}$.

$z_\tau^{\mathtt{AI}}$ inside Ω^*, and thus away from the failure set $\mathcal{F}^{\mathtt{AI}}$, for all time τ. The pessimism of the safety analysis is regulated by restricting the worst-case human behavior to be consistent with the predictive action bound: $a_\tau^{\mathtt{H}} \in \hat{\mathcal{A}}^{\mathtt{H}}(z_\tau^{\mathtt{AI}})$. The construction of these predictive action bounds can once again benefit from the generative AI's predictive power. For example, a large language model can be queried with prompts based on the ODD of the safety analysis and used to sample diverse hypothetical human responses to AI generations (e.g., to simulate antagonistic or goal-driven dialogue (Hong et al., 2023)).

4.3 Posing the Safety-Critical Human–AI Game

We now have all the key mathematical components for a rigorous safety analysis of the human–AI interaction loop. We cast the computation of Ω^* as a zero-sum dynamic game between the AI and a *virtual adversary* that chooses the worst-case realization of the human's behavior allowed by the AI's uncertainty. The game's outcome from any initial information state $z_0^{\mathtt{AI}}$, under optimal play, can be encoded through the *safety value function* (Barron & Ishii, 1989; Tomlin et al., 2000; Lygeros, 2004; Mitchell et al., 2005; Fisac et al., 2015):

$$V(z_0^{\mathtt{AI}}) := \max_{\pi^{\mathtt{AI}}} \min_{\hat{\pi}^{\mathtt{H}}} \left(\min_{t \geq 0} \ell(z_t^{\mathtt{AI}}) \right), \qquad \Omega^* = \left\{ z_0^{\mathtt{AI}} \in \mathcal{Z}^{\mathtt{AI}} : V(z_0^{\mathtt{AI}}) \geq 0 \right\}. \tag{4}$$

Here, $\ell : \mathcal{Z}^{\mathtt{AI}} \to \mathbb{R}$ is a safety margin function that measures the proximity of the HAI system to the failure set and encodes $\mathcal{F}^{\mathtt{AI}}$ as the zero sublevel set $\{z^{\mathtt{AI}} : \ell(z^{\mathtt{AI}}) < 0\}$. If the value $V(z_0^{\mathtt{AI}})$ is negative (i.e., $z_0^{\mathtt{AI}} \notin \Omega^*$), this means that, no matter what the AI agent chooses to do, it cannot avoid eventually entering $\mathcal{F}^{\mathtt{AI}}$ under the worst-case realization of the allowable human actions $a_t^{\mathtt{H}} \in \hat{\mathcal{A}}^{\mathtt{H}}(z_t^{\mathtt{AI}})$ over time. Critically, the game posed in Equation 4 quantifies the best the AI system could ever do to maintain safety—hence, the *maximal* safe set.

The value function defined above satisfies the fixed-point Isaacs equation (Isaacs, 1954) (the game-theoretic counterpart of the Bellman equation) relating the current safety margin ℓ to the minimum-margin-to-go V after one round of play:

$$V(z^{\mathtt{AI}}) = \max_{a^{\mathtt{AI}} \in \mathcal{A}^{\mathtt{AI}}} \min_{a^{\mathtt{H}} \in \hat{\mathcal{A}}^{\mathtt{H}}(z^{\mathtt{AI}})} \underbrace{\min \left\{ \ell(z^{\mathtt{AI}}), \mathop{\mathbb{E}}_{o^{\mathtt{AI}}} \left[V\big(f^{\mathtt{AI}}(z^{\mathtt{AI}}, a^{\mathtt{AI}}, a^{\mathtt{H}}, o^{\mathtt{AI}})\big) \right] \right\}}_{Q(z^{\mathtt{AI}}, a^{\mathtt{AI}}, a^{\mathtt{H}})}. \tag{5}$$

The solution to this zero-sum dynamic programming equation yields a maximin policy pair $(\pi_{\textcircled{\scriptsize D}}^{\mathtt{AI}}, \pi_\dagger^{\mathtt{H}})$ containing the AI's best safety effort $\pi_{\textcircled{\scriptsize D}}^{\mathtt{AI}}$ to maximize the closest future separation from the failure set, and the worst-case human behavior $\pi_\dagger^{\mathtt{H}}$ that would close this distance and, if possible, make it reach zero.[2] The policies $(\pi_{\textcircled{\scriptsize D}}^{\mathtt{AI}}, \pi_\dagger^{\mathtt{H}})$ can be approximately computed through modern learning-based AI methods such as self-supervised learning (Bansal & Tomlin, 2021) or adversarial self-play RL (Silver et al.; Pinto et al., 2017; Hsu et al., 2023). This enables scalable learning from experience and even under partial observability (Hu et al., 2023), and once again leverages the complementary strengths of AI and control systems.

We emphasize that the human behavior encoded by $\pi_\dagger^{\mathtt{H}}$ constitutes a worst-case model (rather than a statistically calibrated one), trained to thwart the AI's best effort to maintain safety but required to conform to the operational design domain. We discuss some important implications of this choice in the conclusion.

In the next section, we discuss how this theoretical human–AI game can be translated into a practical computational procedure enabling AI systems to monitor and enforce safety as they interact with people.

5 Frontier Framework: The Human–AI Safety Filter

As AI technology continues to advance, manually designing or fine-tuning harm prevention strategies with human feedback becomes increasingly untenable (Christiano et al., 2018; Bowman et al., 2022). To break this scalability mismatch, we posit that the same advances driving AI power can be leveraged to *autonomously* identify potential harms and devise proactive strategies that explicitly consider human–AI feedback loops.

[2]The virtual adversary $\pi_\dagger^{\mathtt{H}} : \mathcal{Z}^{\mathtt{AI}} \to \mathcal{A}^{\mathtt{H}}$ exploits the range of (1) plausible internal human states $\hat{z}^{\mathtt{H}} \in \hat{\mathcal{Z}}^{\mathtt{H}}(z^{\mathtt{AI}})$ given the AI's imperfect situational awareness $z^{\mathtt{AI}}$ and (2) ODD-compatible human actions $a^{\mathtt{H}} \in \hat{\mathcal{A}}^{\mathtt{H}}(\hat{z}^{\mathtt{H}})$ given each possible inferred internal state $\hat{z}^{\mathtt{H}}$. Implementations of $\pi_\dagger^{\mathtt{H}}$ may include two-step pipelines ($z^{\mathtt{AI}} \mapsto \hat{z}^{\mathtt{H}} \mapsto a^{\mathtt{H}}$) or implicit end-to-end models ($z^{\mathtt{AI}} \mapsto a^{\mathtt{H}}$).

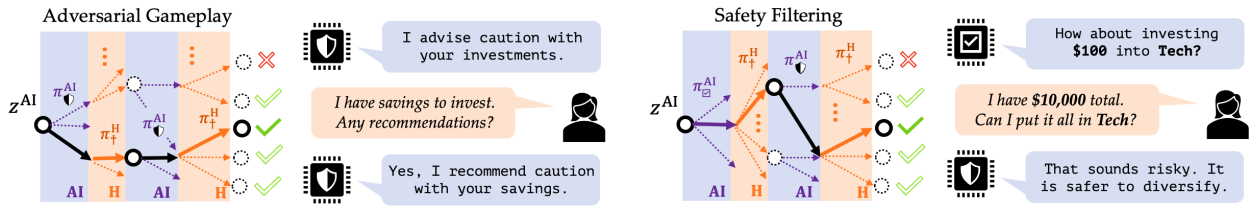


Figure 4: (left) The AI always acts under the safety-critical game policy $(\pi^{\mathtt{AI}}_{\blacklozenge}, \pi^{\mathtt{H}}_{\dagger})$, making it safe but conservative. (right) The filtered AI uses task policy $\pi^{\mathtt{AI}}_{\boxtimes}$ as long as in the *future* it can apply $\pi^{\mathtt{AI}}_{\blacklozenge}$ against $\pi^{\mathtt{H}}_{\dagger}$.

The general formulation in Section 4 enables AI systems to preempt potential pitfalls within a specified ODD, but the resulting policy is *only* concerned with safety. If we were to leave $\pi^{\mathtt{AI}}_{\blacklozenge}$ in control of the AI's entire behavior, as illustrated on the left of Figure 4, it would surely be safe but likely overcautious and unable to provide value to its users. In reality, it is not enough for the AI system to avoid causing failures (if so, we could simply not turn it on), but rather it must do so while assisting its users and performing requested tasks (which may or may not be themselves related to safety). Ideally, we want to *minimally* override the AI's task-driven actions with the safety policy, only intervening at the last possible moment. How can we do this?

The systematic detect-and-avoid functionality we seek closely mirrors the *safety filter* mechanisms established in robotics and control systems, which we reviewed in Section 3. Rather than reinvent a suitable mechanism for human–AI systems, we argue for a frontier framework that builds upon the fundamental principles of safety filtering and extends them to the general interaction problem formalized in Section 4.

Formally, a human–AI safety filter is a tuple $(\pi^{\mathtt{AI}}_{\blacklozenge}, \Delta, \phi)$ containing:

- fallback policy: $\pi^{\mathtt{AI}}_{\blacklozenge} \colon \mathcal{Z}^{\mathtt{AI}} \to \mathcal{A}^{\mathtt{AI}}$, aims only to avoid catastrophic failures, without regard to task performance, and is therefore kept as a last resort.
- safety monitor: $\Delta \colon \mathcal{Z}^{\mathtt{AI}} \times \mathcal{A}^{\mathtt{AI}} \to \mathbb{R}$, checks if the fallback $\pi^{\mathtt{AI}}_{\blacklozenge}$ would still maintain safety *after* a candidate action $a^{\mathtt{AI}}$ is taken from $z^{\mathtt{AI}}$, outputting a positive or negative value following Equation 4.
- intervention scheme: $\phi \colon \mathcal{Z}^{\mathtt{AI}} \times \mathcal{A}^{\mathtt{AI}} \to \mathcal{A}^{\mathtt{AI}}$, permits a candidate action $a^{\mathtt{AI}}$ if if passes the monitoring check and otherwise replaces it with an alternative action that does, for example $\pi^{\mathtt{AI}}_{\blacklozenge}(z^{\mathtt{AI}})$.

This definition can encompass a broad spectrum of potential future supervisory mechanisms (Legg, 2023) and allows us to construct a new central theorem to understand their guarantees and assumptions.

Theorem 1 (General Human–AI Safety Filter) *Consider a human–AI system with AI world model* $f^{AI}(z^{AI}, a^{AI}, a^{H}, o^{AI})$ *and a safety filter* $(\pi^{AI}_{\blacklozenge}, \Delta, \phi)$. *If the AI agent is deployed with an initial internal state* $z^{AI}_0 \in \mathcal{Z}^{AI}$ *deemed safe by the safety monitor under the fallback policy, i.e.,* $\Delta\left(z^{AI}_0, \pi^{AI}_{\blacklozenge}(z^{AI}_0)\right) \geq 0$, *then the interaction under filtered dynamics* $f^{AI}\big(z^{AI}, \phi(z^{AI}, a^{AI}), a^{H}, o^{AI}\big)$ *with any AI task policy* $\pi^{AI}_{\boxtimes} : z^{AI} \mapsto a^{AI}$ *and any realization of human behavior satisfying* $a^{H} \in \hat{\mathcal{A}}^{H}(z^{AI})$ *maintains the safety condition* $\forall t \geq 0, z^{AI} \notin \mathcal{F}^{AI}$.

To date, the concept of a safety filter has only been instantiated for embodied systems with physics-based state spaces (low-dimensional vectors of physical quantities like positions or velocities, governed by well-understood dynamical laws). Here, we are the first to generalize the scope of this mathematical formalism to the much broader context of AI safety. This result lays the theoretical foundations for the algorithmic application of safety filters to *general* human–AI systems, which evolve "latent state spaces" and encode harder to model interactions such as dialogue between a human user and an AI chatbot.

An important aspect of Theorem 1 is that it holds for an arbitrary fallback policy $\pi^{\mathtt{AI}}_{\blacklozenge}$: as long as the safety monitor Δ can accurately predict whether $\pi^{\mathtt{AI}}_{\blacklozenge}$ will succeed at maintaining safety in the future, the intervention scheme ϕ can prevent actions that would lead to a vulnerable state, i.e. a state outside the *fallback-safe* set $\Omega^{\blacklozenge}$. Naturally, if the available fallback policy is not very effective, the filter will be forced to intervene often, restricting the human–AI interactions to remain inside a smaller set $\Omega^{\blacklozenge}$. This is where the safety game from Section 4 comes in.

The Perfect Human–AI Safety Filter. The safety-critical human–AI game we posed in Section 4 implicitly encodes the *least-restrictive* safety filter possible: one that allows maximal freedom to the AI's task policy $\pi^{\mathtt{AI}}_{\boxdot}$ while preempting all future safety failures under the AI's uncertainty. In particular, if we had access to the exact solution to this safety game, such a *perfect* safety filter could be implemented by choosing fallback policy $\pi^{\mathtt{AI}}_{\textcircled{\scriptsize\textbf{!}}}$, safety monitor $\Delta := Q(\cdot, \cdot, \pi^{\mathtt{H}}_{\dagger})$, and switch-type intervention scheme $\phi := \mathbb{1}_{\{\Delta > 0\}} \cdot \pi^{\mathtt{AI}}_{\boxdot} + \mathbb{1}_{\{\Delta \leq 0\}} \cdot \pi^{\mathtt{AI}}_{\textcircled{\scriptsize\textbf{!}}}$.

Algorithmic Human–AI Safety Filtering. We conclude by giving a concrete account of how one could practically instantiate a human–AI safety filter for a generative AI model, visualized in Figure 5. Following the common neural network architecture of generative AI models, let ***base AI model*** (given to us for analysis and safe integration) be comprised of an encoder $\mathcal{E}$ and a decoder $\pi^{\mathtt{AI}}_{\boxdot}$; this decoder is precisely what we have been referring to as the *task policy*, mapping an internal (latent) state $z^{\mathtt{AI}}$ to a proposed output action $a^{\mathtt{AI}}$. The purple block in Figure 5 depicts the ***safety filter*** components: the fallback policy, safety monitor, and intervention scheme. Computationally, adversarial reach-avoid RL can be used to obtain an *approximation* of the optimal fallback policy $\pi^{\mathtt{AI}}_{\textcircled{\scriptsize\textbf{!}}}$ from the safety game in Equation 4. A reliable safety monitor Δ can be implemented by either directly evaluating the learned safety value function at any information state $z^{\mathtt{AI}}$ (safety critic) or by simulating a family of pessimistic interaction scenarios by querying the learned *virtual adversary* $\pi^{\mathtt{H}}_{\dagger}$. In turn, the intervention scheme can range from a simple binary switch (at each time, apply $\pi^{\mathtt{AI}}_{\boxdot}$ if deemed safe, else apply $\pi^{\mathtt{AI}}_{\textcircled{\scriptsize\textbf{!}}}$) to a more sophisticated override (e.g., find a minimally disruptive deviation from $\pi^{\mathtt{AI}}_{\boxdot}$ that is deemed safe).

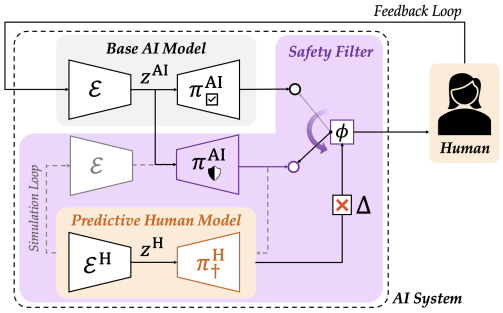


Figure 5: **Human–AI Safety Filter**. The base AI model encodes the AI's observations into its latent state $z^{\mathtt{AI}}$ which is used as input for its task policy ($\pi^{\mathtt{AI}}_{\boxdot}$). A safety filter includes a learned AI safety strategy $\pi^{\mathtt{AI}}_{\textcircled{\scriptsize\textbf{!}}}$, a safety monitor Δ that predicts safety risks, and a predictive human model containing a virtual adversary $\pi^{\mathtt{H}}_{\dagger}$ that generates pessimistic predictions of human interaction. Based on Δ, the AI's outputs to the human are filtered by the intervention scheme ϕ, and modified to guarantee safety.

Even though the components of the safety filter would be approximate by their learning-based nature, the scheme can be leveraged in combination with modern statistical generalization theory, such as PAC-Bayes theory (McAllester, 2003; Majumdar et al., 2020), adversarial conformal prediction (Gibbs & Candes, 2021; Bastani et al., 2022), and scenario optimization (Schildbach et al., 2014; Lin & Bansal, 2023), to maintain a high-confidence guarantee that the AI system will robustly enforce the satisfaction of the human's critical needs throughout the interaction for *all* human behaviors allowed by the operational assumptions. We emphasize that a key strength of this safety framework is that it naturally scales with the rapidly advancing *capability* of modern AI systems: as future generations of language models, vision-language systems, and general AI agents become ever stronger, so will the safety assurances that can be provided through the proposed techniques and system architecture.

6 Conclusion

In this paper, we aim to inspire the genesis of a new human–AI safety research community. We take concrete steps towards this by identifying a fundamental synergy between the principled safety formalism offered by control theory and the general representations learned by internet-trained AI systems. By combining lessons from control and AI, we propose a technical roadmap to guide research efforts towards a safety framework that can reliably anticipate and avoid potential hazards emerging during interaction. We propose a frontier framework called the *human–AI safety filter*, wherein an AI system's task policy is systematically monitored and minimally overriden by a safety policy synthesized via safety-critical adversarial self-play.

Broader Impact Statement

We expect that the proposed interdisciplinary safety framework will help catalyze a much needed rapprochement between the AI and control systems communities to develop rigorous safety assurances for dynamic human–AI feedback loops. A significant positive impact may come in the form of the first practical safety

frameworks that can not only keep up with the rapid advances in AI capabilities but actively benefit from them to provide stronger guarantees, ushering in a new generation of advanced AI systems that can be trusted *because* of their sophistication, and not in spite of it.

On the other hand, we also highlight possible pitfalls of our proposed human–AI safety framework. The approximate nature of learning-based generative AI makes it extremely challenging to provide a clear-cut delineation of uncertainty, which will likely limit us to statistical assurances in the foreseeable future. These fall short of the stronger *if–then* certificates that we can aspire to in other engineering domains: hard guarantees establishing that, as long as the system's operational assumptions are met, catastrophic failures are categorically impossible (i.e., a failure can only result from an explicit assumption being violated). Even high-confidence statistical assurances can leave human-centered AI systems open to black swan events with extremely low probability but potentially dramatic consequences.

There is a risk that the improved treatment of human–AI feedback loops developed through the proposed agenda could be repurposed and misused by malicious or reckless actors to construct AI systems that exploit interaction dynamics against the interest of their users, for example by seeking to manipulate their decisions. Even with today's relatively myopic fine-tuning approaches, we see a worrying emergence of unintended (e.g., sycophantic) AI outputs as the system learns to secure positive user responses. Future systems equipped with long-horizon reasoning but *without* a proper safety framework could conceivably seek long-term interaction outcomes serving a third party's agenda at the expense of their users' needs.

We nonetheless remain cautiously optimistic: First, human–AI safety filtering does not require teasing apart the likelihood of various conceivable human behaviors in a given context. Rather, safety-directed predictions robustly consider the set of all such plausible behaviors without distinction, making them harder to exploit for manipulation purposes. Second, the need to consider large prediction sets containing both likely and unlikely outcomes aligns well with the inclusion of underrepresented individual behaviors that do not conform to dominant patterns in the training datasets. Finally, provided that future AI systems are deployed with a cyber-secure dynamical safety mechanism that cannot be removed or altered by unauthorized parties, such a framework would help detect and mitigate emergent and intentional misalignment. Naturally, this will require a process of standardization and regulatory oversight; the first step, however, must be to establish *what assurances are possible*. Ultimately, we expect that technical advances in human–AI safety will inform the conversation between technologists, policymakers, political leaders, and the public at large. A timely conversation that, fortunately, is already ongoing.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anayo K. Akametalu, Shromona Ghosh, Jaime F. Fisac, Vicenc Rubies-Royo, and Claire J. Tomlin. A minimum discounted reward Hamilton–Jacobi formulation for computing reachable sets. 69(2):1097–1103. ISSN 1558-2523. doi: 10.1109/TAC.2023.3327159. URL `https://ieeexplore.ieee.org/document/10294099`.

Ross E Allen, Ashley A Clark, Joseph A Starek, and Marco Pavone. A machine learning approach for real-time reachability analysis. In *2014 IEEE/RSJ international conference on intelligent robots and systems*, pp. 2202–2208. IEEE, 2014.

Matthias Althoff and John M Dolan. Online verification of automated road vehicles using reachability analysis. *IEEE Transactions on Robotics*, 30(4):903–918, 2014.

Matthias Althoff, Colas Le Guernic, and Bruce H Krogh. Reachable set computation for uncertain time-varying linear systems. In *Proceedings of the 14th international conference on Hybrid systems: computation and control*, pp. 93–102, 2011.

Aaron D Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *2019 18th European control conference (ECC)*, pp. 3420–3431. IEEE, 2019.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Andrea Bajcsy, Somil Bansal, Ellis Ratner, Claire J Tomlin, and Anca D Dragan. A robust control framework for human motion prediction. *Robotics and Automation Letters*, 2020.

Somil Bansal and Claire J Tomlin. DeepReach: A deep learning approach to high-dimensional reachability. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J Tomlin. Hamilton-Jacobi Reachability: A brief overview and recent advances. In *IEEE Conference on Decision and Control (CDC)*, 2017.

EN Barron and H Ishii. The bellman equation for minimizing the maximum cost. *NONLINEAR ANAL. THEORY METHODS APPLIC.*, 13(9):1067–1090, 1989.

Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

Osbert Bastani. Safe reinforcement learning with nonlinear dynamics via model predictive shielding. In *2021 American control conference (ACC)*, pp. 3488–3494. IEEE, 2021.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35:29362–29373, 2022.

Luca Bergamini, Yawei Ye, Oliver Scheel, Long Chen, Chih Hu, Luca Del Pero, Błażej Osiński, Hugo Grimmett, and Peter Ondruska. Simnet: Learning reactive self-driving simulations from real-world observations. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5119–5125. IEEE, 2021.

Roger V. Bostelman, Joseph A. Falco, Marek Franaszek, and Kamel S. Saidi. Performance assessment framework for robotic systems. Technical report, National Institute of Standards and Technology, 2018.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

Yuxiao Chen, Andrew Singletary, and Aaron D Ames. Guaranteed obstacle avoidance for multi-robot operations with limited actuation: A control barrier function approach. *IEEE Control Systems Letters*, 5(1): 127–132, 2020.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *International Conference on Learning Representations*, 2024.

Jérôme Darbon, Gabriel P Langlois, and Tingwei Meng. Overcoming the curse of dimensionality for some hamilton–jacobi partial differential equations via neural network architectures. *Research in the Mathematical Sciences*, 7:1–50, 2020.

Charles Dawson, Zengyi Qin, Sicun Gao, and Chuchu Fan. Safe nonlinear control using robust neural lyapunov-barrier functions. In *Conference on Robot Learning*, pp. 1724–1735. PMLR, 2022.

Charles Dawson, Sicun Gao, and Chuchu Fan. Safe control with learned certificates: A survey of neural lyapunov, barrier, and contraction methods for robotics and control. *IEEE Transactions on Robotics*, 2023.

Roel Dobbe, Patricia Hidalgo-Gonzalez, Stavros Karagiannopoulos, Rodrigo Henriquez-Auba, Gabriela Hug, Duncan S Callaway, and Claire J Tomlin. Learning to control in power systems: Design and analysis guidelines for concrete safety problems. *Electric Power Systems Research*, 189:106615, 2020.

Katherine Driggs-Campbell, Roy Dong, and Ruzena Bajcsy. Robust, informative human-in-the-loop predictions via empirical reachable sets. *IEEE Transactions on Intelligent Vehicles*, 3(3):300–309, 2018.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *arXiv preprint arXiv:2302.00111*, 2023.

J. Fisac, M. Chen, C. J. Tomlin, and S. Sastry. Reach-avoid problems with time-varying dynamics, targets and constraints. In *HSCC*, 2015.

J. F. Fisac, N. F. Lugovoy, V. Rubies-Royo, S. Ghosh, and C. J. Tomlin. Bridging Hamilton-Jacobi safety analysis and reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8550–8556, 2019. doi: 10.1109/ICRA.2019.8794107.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. Aligning llm agents by learning latent preference from user edits. *arXiv preprint arXiv:2404.15269*, 2024.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672. Curran Associates, Inc., 2021.

Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. Drocc: Deep robust one-class classification. In *International conference on machine learning*, pp. 3711–3721. PMLR, 2020.

Lin Guan, Yifan Zhou, Denis Liu, Yantian Zha, Heni Ben Amor, and Subbarao Kambhampati. " task success" is not enough: Investigating the use of video-language models as behavior critics for catching undesirable agent behaviors. *arXiv preprint arXiv:2402.04210*, 2024.

David Ha and Jürgen Schmidhuber. World models. *NIPS*, 2018.

Sami Haddadin, Simon Haddadin, Augusto Khoury, Tim Rokahr, Sven Parusel, Rainer Burgkart, Antonio Bicchi, and Alin Albu-Schäffer. On making robots understand safety: Embedding injury knowledge into control. *The International Journal of Robotics Research*, 31(13):1578–1602, 2012.

Alex Heath. All the news from openai's first developer conference. https://www.theverge.com/2023/11/6/23948619/openai-chatgpt-devday-developer-conference-news, 2023. Accessed: 2024-01-23.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Lukas Hewing, Kim P Wabersich, Marcel Menner, and Melanie N Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:269–296, 2020.

Coryanne Hicks. I pitted chatgpt against a real financial advisor to help me save for retirement—and the winner is clear. https://fortune.com/recommends/investing/chatgpt-vs-real-financial-advisor-to-plan-retirement-which-is-better/, 2023. Accessed: 2024-01-12.

Joey Hong, Sergey Levine, and Anca Dragan. Zero-shot goal-directed dialogue via rl on imagined conversations. *arXiv preprint arXiv:2311.05584*, 2023.

Kai-Chieh Hsu, Duy Phuong Nguyen, and Jaime Fernàndez Fisac. ISAACS: Iterative Soft Adversarial Actor-Critic for Safety. *Learning for Dynamics and Control Conference*, pp. 90–103, 2023.

Kai-Chieh Hsu, Haimin Hu, and Jaime Fernández Fisac. The safety filter: A unified view of safety-critical control in autonomous systems. *Annual Reviewsof Control, Robotics, and Autonomous Systems*, 2024.

Haimin Hu, Zixu Zhang, Kensuke Nakamura, Andrea Bajcsy, and Jaime F Fisac. Deception game: Closing the safety-learning loop in interactive robot autonomy. *Conference on Robot Learning*, 2023.

International Civil Aviation Organization ICAO. The world of air transport in 2019. https://www.icao.int/annual-report-2019/Pages/the-world-of-air-transport-in-2019.aspx#:~:text=According%20to%20ICAO's%20preliminary%20compilation,a%201.7%20per%20cent%20increase, 2019. Accessed: 2024-05-28.

Rufus Isaacs. *Differential games I*. RAND Corporation, 1954.

ISO 15066. Robots and robotic devices – Collaborative robots. Standard, International Organization for Standardization, 2016.

ISO 22737:2021. Intelligent transport systems. Standard, International Organization for Standardization, 2021.

Boris Ivanovic, Edward Schmerling, Karen Leung, and Marco Pavone. Generative modeling of multimodal multi-human behavior. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3088–3095. IEEE, 2018.

Mykel J Kochenderfer, Jessica E Holland, and James P Chryssanthacopoulos. Next generation airborne collision avoidance system. *Lincoln Laboratory Journal*, 19(1):17–33, 2012.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Shane Legg. System two safety. The Alignment Workshop, 2023. URL https://www.alignment-workshop.com/nola-talks/shane-legg-system-two-safety.

Anjian Li, Liting Sun, Wei Zhan, Masayoshi Tomizuka, and Mo Chen. Prediction-based reachability for collision avoidance in autonomous driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7908–7914. IEEE, 2021a.

Jiacheng Li, Qingchen Liu, Wanxin Jin, Jiahu Qin, and Sandra Hirche. Robust safe learning and control in an unknown environment: An uncertainty-separated control barrier function approach. *IEEE Robotics and Automation Letters*, 2023.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. A systematic investigation of commonsense knowledge in large language models. *Conference on Empirical Methods in Natural Language Processing*, 2021b.

Albert Lin and Somil Bansal. Generating formal safety assurances for high-dimensional reachability. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10525–10531. IEEE, 2023.

Stefan B Liu, Hendrik Roehm, Christian Heinzemann, Ingo Lütkebohle, Jens Oehlerking, and Matthias Althoff. Provably safe motion of mobile robots in human environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1351–1357. IEEE, 2017.

Matt Luckcuck, Marie Farrell, Louise A Dennis, Clare Dixon, and Michael Fisher. Formal specification and verification of autonomous robotic systems: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–41, 2019.

John Lygeros. On reachability and minimum cost optimal control. *Automatica*, 40(6):917–927, 2004.

Anirudha Majumdar, Alec Farid, and Anoopkumar Sonar. Pac-bayes control: Learning policies that provably generalize to novel environments, 2020.

Tommaso Mannucci, Erik-Jan van Kampen, Cornelis De Visser, and Qiping Chu. Safe exploration algorithms for reinforcement learning controllers. *IEEE transactions on neural networks and learning systems*, 29(4): 1069–1081, 2017.

K. Margellos and J. Lygeros. Hamilton-Jacobi Formulation for Reach–Avoid Differential Games. *IEEE Trans. on Automatic Control*, 56(8):1849–1861, 2011.

David McAllester. Simplified pac-bayesian margin bounds. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pp. 203–215. Springer, 2003.

Fundamental AI Research Diplomacy Team (FAIR) Meta, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

Ian Mitchell, Alex Bayen, and Claire J. Tomlin. A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control (TAC)*, 50(7):947–957, 2005.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Dan Hendrycks, and David Wagner. Can llms follow simple rules? *arXiv preprint arXiv:2311.04235*, 2023.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Kensuke Nakamura and Somil Bansal. Online update of safety assurances using confidence-based predictions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12765–12771. IEEE, 2023.

Jack Nicas, Natalie Kitroeff, David Gelles, and James Glanz. Boeing built deadly assumptions into 737 max, blind to a late design change. `https://www.nytimes.com/2019/06/01/business/boeing-737-max-crash.html`, 2019. Accessed: 2024-01-24.

Dan Noyes. New video of bay bridge 8-car crash shows tesla abruptly braking in 'self-driving' mode. https://abc7news.com/tesla-sf-bay-bridge-crash-8-car-self-driving-video/12686428/, 2023. Accessed: 2024-01-24.

SAE On-Road Automated Driving Committee. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles, 2021. URL https://www.sae.org/content/j3016_202104.

OpenAI. Our approach to alignment research. https://openai.com/blog/our-approach-to-alignment-research, 2022a. Accessed: 2024-01-23.

OpenAI. Dall-e now available without waitlist. https://openai.com/blog/dall-e-now-available-without-waitlist, 2022b. Accessed: 2024-01-23.

OpenAI. DALL · E 2 preview: Risks and limitations. https://github.com/openai/dalle-2-preview/blob/main/system-card_04062022.md, 2022c. Accessed: 2024-01-23.

Pedro A. Ortega, Vishal Maini, and DeepMind Safety Team. Building safe artificial intelligence: specification, robustness, and assurance. https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1, 2018. Accessed: 2024-01-23.

Piotr F. Orzechowski, Kun Li, and Martin Lauer. Towards Responsibility-Sensitive Safety of Automated Vehicles with Reachable Set Analysis. In *2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE)*, pp. 1–6. IEEE. ISBN 978-1-72810-142-2. doi: 10.1109/ICCVE45908.2019.8965069. URL https://ieeexplore.ieee.org/document/8965069/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,*, 2022.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Zengyi Qin, Kaiqing Zhang, Yuxiao Chen, Jingkai Chen, and Chuchu Fan. Learning safe multi-agent control with decentralized neural barrier certificates. *International Conference on Learning Representations*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Vicenç Rubies-Royo, David Fridovich-Keil, Sylvia Herbert, and Claire J Tomlin. A classification-based approach for approximate reachability. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7697–7704. IEEE, 2019.

Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

Georg Schildbach, Lorenzo Fagiano, Christoph Frei, and Manfred Morari. The scenario approach for stochastic model predictive control with bounds on closed-loop constraint violations. *Automatica*, 50(12):3009–3018, 2014.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pp. 146–157. Springer, 2017.

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8579–8590, 2023.

Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. Preferences implicit in the state of the world. *International Conference on Learning Representations*, 2019.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. 362(6419):1140–1144. doi: 10.1126/science.aar6404. URL `https://www.science.org/doi/full/10.1126/science.aar6404`.

Sumeet Singh, Anirudha Majumdar, Jean-Jacques Slotine, and Marco Pavone. Robust online motion planning via contraction theory and convex optimization. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5883–5890. IEEE, 2017.

Simon Suo, Sebastian Regalado, Sergio Casas, and Raquel Urtasun. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10400–10409, 2021.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Ran Tian, Liting Sun, Andrea Bajcsy, Masayoshi Tomizuka, and Anca D Dragan. Safety assurances for human-robot interaction via confidence-aware game-theoretic human models. In *IEEE International Conference on Robotics and Automation*, pp. 11229–11235, 2022.

C.J. Tomlin, J. Lygeros, and S. Shankar Sastry. A game theoretic approach to controller design for hybrid systems. 88(7):949–970, 2000. ISSN 0018-9219, 1558-2256. doi: 10.1109/5.871303. URL `http://ieeexplore.ieee.org/document/871303/`.

Claire Tomlin, George J Pappas, and Shankar Sastry. Conflict resolution for air traffic management: A study in multiagent hybrid systems. *IEEE Transactions on automatic control*, 43(4):509–521, 1998.

The European Union. A european approach to artificial intelligence. `https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence`, 2021. Accessed: 2024-01-23.

Michael Vitus and Claire Tomlin. Hierarchical, hybrid framework for collision avoidance algorithms in the national airspace. In *AIAA Guidance, Navigation and Control Conference and Exhibit*, pp. 6970, 2008.

Kim P Wabersich, Andrew J Taylor, Jason J Choi, Koushil Sreenath, Claire J Tomlin, Aaron D Ames, and Melanie N Zeilinger. Data-driven safety filters: Hamilton-jacobi reachability, control barrier functions, and predictive methods for uncertain systems. *IEEE Control Systems Magazine*, 43(5):137–177, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*, 2023.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

United States White House. Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence, 2023.

Norbert Wiener. Some moral and technical consequences of automation. 131(3410):1355–1358. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.131.3410.1355. URL https://science.sciencemag.org/content/131/3410/1355.

Chloe Xiang. "he would still be here": Man dies by suicide after talking with ai chatbot, widow says. https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says, 2023. Accessed: 2024-01-23.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2950–2968, 2021.

Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Transactions On Graphics (TOG)*, 42(4):1–14, 2023.

Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3560–3566. IEEE, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.