Position: Safe AI Should be Resistant and Resilient in an Evolving World

Anonymous Authors[1]

Abstract

In this position paper, we address the persistent gap between rapidly growing AI capabilities and lagging safety progress. Existing paradigms divide into "Make AI Safe", which applies post-hoc alignment and guardrails but remains brittle and reactive, and "Make Safe AI", which emphasizes intrinsic safety but struggles to address unforeseen risks in open-ended environments. We therefore propose *safe-by-coevolution* as a new formulation of the "Make Safe AI" paradigm, inspired by biological immunity, in which safety becomes a dynamic, adversarial, and ongoing learning process. To operationalize this vision, we introduce R^{2}AI—*Resistant and Resilient AI*—as a practical framework that unites resistance against known threats with resilience to unforeseen risks. R^{2}AI integrates *fast and slow safe models*, adversarial simulation and verification through a *safety wind tunnel*, and continual feedback loops that guide safety and capability to coevolve. We argue that this framework offers a scalable and proactive path to maintain continual safety in dynamic environments, addressing both near-term vulnerabilities and long-term existential risks as AI advances toward AGI and ASI.

1. Introduction

Recent years have witnessed rapid developments and huge breakthroughs in AI, leading to its integration into everyday life and establishing it as a foundational infrastructure in society (Van Der Vlist et al., 2024). As AI systems are increasingly deployed in safety-critical domains (*e.g.,* scientific research (Jumper et al., 2021; Zhang et al., 2023; Novikov et al., 2025), autonomous driving (Wang et al., 2021; Rowe et al., 2024), healthcare (Panayides et al., 2020; Bekbolatova et al., 2024), law (Lai et al., 2024)), the risks posed by unsafe or unreliable outputs have become more pronounced. In such settings, failures can result in severe, even catastrophic, consequences. Beyond these near-term concerns, the continued advancement toward highly autonomous and superhuman-level AI raises long-term existential risks (Dalrymple et al., 2024; Bengio et al., 2025a;b; Kulveit et al., 2025; Clymer et al., 2025; Shanghai AI Lab, 2025a). As capabilities scale, so does the difficulty of aligning, controlling, and governing these systems, thus potentially leading to scenarios with irreversible societal or civilizational impacts (Bengio et al., 2025c; Shanghai AI Lab & Concordia AI, 2025).

Despite escalating risks, safety progress has lagged far behind capability growth. As shown in Figure 1a, evaluations show a consistent pattern: leading AI models worldwide—such as GPT-5 (OpenAI, 2025), Claude 4 (Anthropic, 2025), and Gemini 3 Pro (Google DeepMind, 2025)—demonstrate significantly higher capability scores than safety scores. This imbalance reveals a structural problem: current safety approaches are reactive, fragmented, and incapable of scaling with capability. To capture this tension, Yang et al. (2024) proposed the AI-45° Law: safety and capability must coevolve along a 45° diagonal trajectory. Temporary deviations are tolerable, but persistent dips below the 45° line increase the risk of catastrophic misalignment, while rising above it may unnecessarily stall innovation. We further define two thresholds: *yellow lines* serve as early warnings when capability begins to outpace safety; *red lines* denote irreversible, catastrophic risks that must never be crossed (IDAIS, 2024; 2025).

Current safety research can be broadly categorized into two paradigms. The dominant "Make AI Safe" paradigm seeks to improve safety after model development, typically through alignment fine-tuning (*e.g.,* RLHF (Christiano et al., 2017), RLAIF (Bai et al., 2022)), red teaming (Perez et al., 2022; Ganguli et al., 2022; Pavlova et al., 2024), and guardrail (Bai et al., 2022; Rajpal, 2023; Oh et al., 2024). While effective in mitigating known risks, these methods are often reactive, brittle, expensive, and struggle to address unknown or emerging risks. In contrast, the "Make Safe AI"

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Figure 1a is reproduced from data available at https://aiben.ch. Figures 1b and 1c are adapted from Figure 1 in Yang et al. (2024).

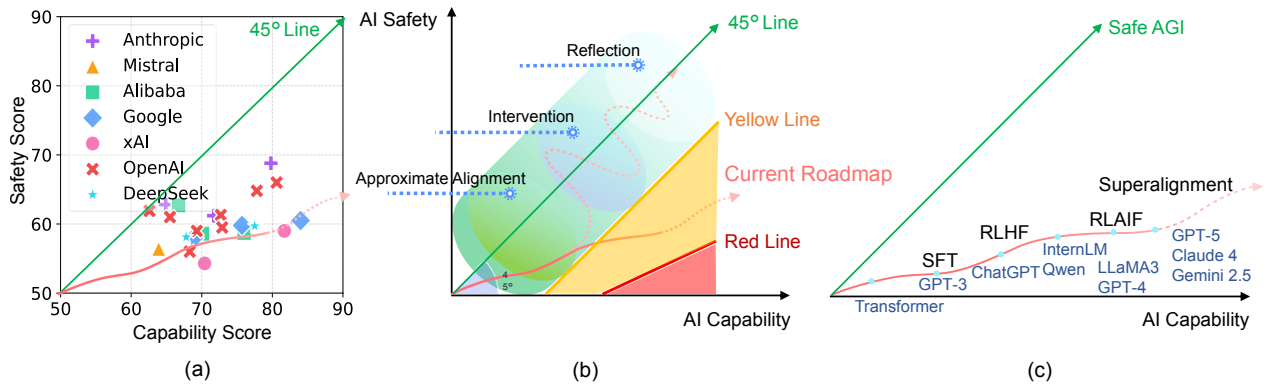

Figure 1. The AI-45° Law (Yang et al., 2024): coevolving capability with safety.[1] (a) Empirical distribution of leading foundation models, showing a widening gap between capability scores and safety scores across major labs. (b) Conceptual safety–capability plane comparing the current roadmap (pink) with the yellow, red, and 45° trajectories toward safe AGI, emphasizing transitions from approximate alignment to reflection. (c) Historical timeline of frontier models, from Transformer (Vaswani et al., 2017) to GPT-5 (OpenAI, 2025), Claude-4 (Anthropic, 2025), and Gemini-2.5 (Comanici et al., 2025), illustrating the divergence between capability scaling and current alignment methods (*e.g.,* SFT (Ouyang et al., 2022), RLHF (Christiano et al., 2017), RLAIF (Bai et al., 2022)), and the need for a coevolutionary path to Safe AGI.

paradigm emphasizes intrinsic safety, designing systems to be safe by construction. Prominent directions include formal guarantees (Szegedy, 2020; Dalrymple et al., 2024) and Scientist AI (Bengio et al., 2025a). Yet even these approaches often fall short in open-ended environments where novel risks cannot be fully anticipated.

To achieve *scalable safety* in an evolving world, we must rethink what "Make Safe AI" entails. We argue that its core principle should be *coevolution*: safety must not be treated as a constraint or one-time guarantee, but as a continuous, adaptive capability that evolves alongside intelligence in uncertain, dynamic environments. We therefore propose *safe-by-coevolution* as a new formulation for "Make Safe AI", inspired by biological immunity (Cooper & Alder, 2006; Müller et al., 2018; Papkou et al., 2019), in which safety becomes a dynamic, adversarial, and ongoing learning process. By embedding coevolutionary mechanisms into the AI lifecycle, systems can remain safe through sustained interaction with real and simulated environments. Just as human immunity develops through continual exposure to pathogens (Flajnik & Kasahara, 2010; Nourmohammad et al., 2016; Buckingham & Ashby, 2022), AI must develop safety through ongoing interaction with its environment. Without such an "immune system", advanced AI risks becoming powerful yet dangerously brittle, and unlike humans, a single catastrophic failure could be irreversible.

Safe-by-coevolution advances a proactive path for safety evolution. It is structured around three iterative steps: 1) *Near-term safety guarantee*: ensure that an AI system at time t_0 is verifiably within a defined safety margin; 2) *Safe iterative step*: for any system already safe, design coevolutionary mechanisms—adversarial interactions, feedback loops, and continuous updates—to guide each upgrade back within that margin; 3) *Continual safety by induction*: repeat this loop so that safety evolves in sync with capability. Unlike reactive patching, this approach integrates safety into the developmental process. To address unforeseen risks (*e.g.,* paradigm shifts or red-line events), it further incorporates a *reset-and-recover* mechanism: halting unsafe systems, redefining safety margins, and establishing new verified checkpoints to sustain coevolution.

To realize this vision, we introduce R^{2}AI—*Resistant and Resilient AI*—as a practical framework for safe-by-coevolution. R^{2}AI unifies *resistance* and *resilience* as the two foundational and complementary dimensions of intrinsic safety: resistance captures robustness against known threats, while resilience emphasizes recovery and adaptation under unforeseen risks. Specifically, R^{2}AI comprises four interacting components: (i) *fast safe models* for real-time response, (ii) *slow safe models* for verification and reasoning, (iii) a *safety wind tunnel* that simulates adversarial attacks and validation loops, and (iv) an *external environment* for interacting with diverse, realistic scenarios. Through adversarial and cooperative dynamics, these components coevolve to embed safety as a learned and adaptive property. Over time, slow mechanisms become internalized into fast, intuitive safeguards, thereby lowering the cost of compliance and enabling scalable, intrinsic safety even at the frontier of AGI (Goertzel, 2014; Lake et al., 2017; Baum, 2017; Bubeck et al., 2023; Morris et al., 2024; Raman et al., 2025).

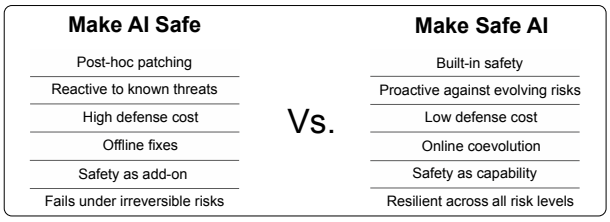

Figure 2. Conceptual contrast between "Make AI Safe" and "Make Safe AI".

Our position. We propose R^{2}AI—a framework uniting resistance and resilience—as a scalable and intrinsically adaptive approach to AI safety. Grounded in *safe-by-coevolution*, it reconceives safety as a continual learning process rather than a static constraint, enabling systems to withstand known threats, adapt to unforeseen risks, and evolve in step with capability. This redefinition offers a generalizable alternative to brittle alignment or top-down control, providing a proactive path to sustain safety across dynamic environments and future ASI (Nick, 2014; Kim et al., 2024; Hendrycks et al., 2025).

Structure of this paper. The remainder of the paper is organized as follows. Section 2 reconsiders the paradigm of "Make Safe AI" and introduces *resistance* and *resilience* as its foundational properties. Section 3 formalizes the *safe-by-coevolution* principle, establishing its theoretical foundation and operational steps. Section 4 presents the R^{2}AI framework, detailing its core components, mechanisms, and continual learning strategies. Section 5 discusses the implications, applications, and societal impacts of R^{2}AI, highlighting its relevance to both near-term safety challenges and long-term existential risks.

2. Rethinking "Make Safe AI"

2.1. Alternative Views

The contrast between "Make AI Safe" and "Make Safe AI", as shown in Figure 2, underscores a fundamental shift in perspective. While "Make AI Safe" relies on post-hoc fixes, reactive defenses, and costly patching that falter under irreversible risks, "Make Safe AI" envisions safety as a built-in, proactive, and evolving capability. This transition requires rethinking safety not as an external add-on, but as an intrinsic property that coevolves with intelligence.

Existing work toward "Make Safe AI" has made important progress—ranging from formal guarantees (Seshia et al., 2022; Dalrymple et al., 2024) to constrained design choices such as Scientist AI (Bengio et al., 2025a) and Tool AI (Karnofsky, 2024). Yet these approaches struggle in open-ended, non-stationary environments where novel objectives, adversarial pressures, and distributional shifts are inevitable.

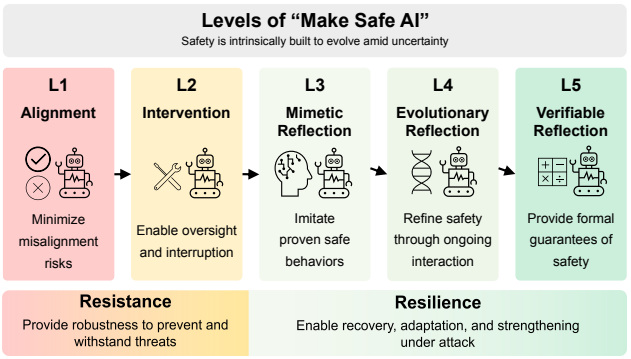

Figure 3. Five levels of "Make Safe AI", which progressively embed safety as an intrinsic and evolving capability.

We argue that the foundation of "Make Safe AI" must rest on two complementary properties, inspired by ecological systems where long-lived organisms survive under continual stress (Holling et al., 1973; Levin, 1998; Gunderson, 2000; Walker et al., 2004): *resistance*, the capacity to withstand and mitigate known threats, and *resilience*, the capacity to recover, adapt, and improve in the face of unforeseen disturbances. Unlike static safeguards, these properties are endogenous, enabling systems to maintain integrity across dynamic and uncertain environments.

Building on this foundation, we propose *safe-by-coevolution* as a new formulation of the "Make Safe AI" paradigm. Inspired by biological immunity (Cooper & Alder, 2006; Müller et al., 2018; Papkou et al., 2019), this principle reconceives safety as a dynamic, adversarial, and ongoing learning process. Rather than attaching fixed safeguards to capable systems, safety itself must scale with capability—reflexively, adaptively, and proactively. This redefinition is essential for ensuring that AI systems preserve both functional integrity and ethical alignment in real-world complexity.

2.2. Levels of "Make Safe AI"

Building on this redefinition of "Make Safe AI", we can further structure its progression into a hierarchy of safety levels. As illustrated in Figure 3, safety is not a binary attribute but an evolving spectrum, intrinsically built to adapt amid uncertainty. This perspective highlights how safety matures from basic alignment toward fully verified, self-evolving guarantees. To instantiate this view, we propose a five-level spectrum that extends the causal ladder of trustworthy AI (Yang et al., 2024). This spectrum reflects increasing degrees of adaptivity, autonomy, and assurance in dynamic environments, capturing the progression from approximate alignment to formal, verifiable safety.

At the foundational layers, *resistance* anchors safety by providing robustness against known risks: *L1 Alignment*

minimizes misalignment through approximate tuning, and *L2 Intervention* ensures oversight and the ability to halt unsafe behavior. Building upward, *resilience* enables adaptive safety beyond reactive correction: *L3 Mimetic Reflection* introduces internal reflection by imitating proven safe behaviors to anticipate risks, *L4 Evolutionary Reflection* advances this reflection into continual co-adaptation with environments, and *L5 Verifiable Reflection* culminates in formalized reflection, where provable guarantees sustain resilience even under uncertainty. Specifically,

- **L1: Alignment.** Safety is achieved through approximate alignment (Yang et al., 2024), typically via supervised fine-tuning, direct preference optimization (Rafailov et al., 2023; Meng et al., 2024; Wu et al., 2024b), reinforcement learning from human feedback (Ouyang et al., 2022; Bai et al., 2022; Shao et al., 2024), knowledge editing (Meng et al., 2022; Fang et al., 2025; Jiang et al., 2025), or activation steering (Arditi et al., 2024; Panickssery et al., 2023). While practical, such alignment is static and correlation-based, providing robustness against known risks but requiring continual updates to withstand new tasks or adversarial strategies (Perez et al., 2023; Zou et al., 2023a; Wei et al., 2023; Yi et al., 2024; Ji et al., 2024).

- **L2: Intervention.** Safety is treated as a control problem (Hendrycks et al., 2021), where systems monitor outputs and intervene when thresholds are violated (Orseau & Armstrong, 2016; Zou et al., 2024; Xu et al., 2024), guided by explicit feedback (Bengio et al., 2025c; Zhu et al., 2025). This level provides oversight and interruption, offering robustness through reactive correction (Ganguli et al., 2022). In addition, advances in mechanistic interpretability (Sharkey et al., 2025) provide tools to identify and intervene on unsafe internal circuits or representations before they manifest in outputs (Nanda et al., 2023; Conmy et al., 2023; Bereska & Gavves, 2024). However, the overall effectiveness depends on timely and reliable feedback signals (Leike et al., 2018; Lin et al., 2021; Terekhov et al., 2025).

- **L3: Mimetic Reflection.** At this level, the system engages in *reflection by imitation*, developing internal reasoning capabilities (Shinn et al., 2023b; Madaan et al., 2023; Guan et al., 2024a; Shanghai AI Lab, 2025b; Zhang et al., 2025a; Yang et al., 2025b; Zhang et al., 2025b). It can perform counterfactual reasoning, simulate outcomes, and anticipate risks by internalizing proven safe behaviors (Dai et al., 2023; Reddy Chirra et al., 2024). This marks a shift from externally imposed oversight to internalized safety reasoning, enabling anticipatory resilience and reducing dependence on continuous supervision.

- **L4: Evolutionary Reflection.** Reflection becomes *evolutionary*: safety mechanisms themselves adapt through continual interaction and coevolution with capabilities and environments (Pan et al., 2025; Cai et al., 2025). Safety thus becomes an agentic property (Wang et al., 2025a)—self-directed, adaptive, and scalable to complex or unforeseen challenges—enabling recovery and strengthening under attack.

- **L5: Verifiable Reflection.** Reflection reaches its most advanced form: *formalized reflection*, where safety reasoning is anchored in mathematical verification (Dalrymple et al., 2024). Systems can not only reflect on possible risks but also prove the correctness of safety guarantees under uncertainty (Vassev, 2016; Bengio et al., 2025a). This integration of formal specification with learning dynamics provides the strongest form of resilient assurance, sustaining trust even in open-ended environments.

Together, these five levels extend the causal ladder of trustworthy AI (Yang et al., 2024) into a coevolving safety framework. This layered progression—from externally imposed safeguards to internalized, self-evolving, and verifiable safety—outlines a roadmap for *safe-by-coevolution*: a reformulation of "Make Safe AI" in which safety is conceived as an intrinsic, reflexive capability that scales alongside intelligence.

3. Safe-by-Coevolution

In this section, we formally introduce *safe-by-coevolution*, a new formulation of "Make Safe AI" that reframes safety as an intrinsic capability evolving alongside intelligence. Rather than relying on reactive defenses (Ganguli et al., 2022; Bai et al., 2022; Zou et al., 2024; Guan et al., 2024a) or externally imposed constraints (Lee et al., 2024; Guan et al., 2024b), this approach envisions systems that sustain safety through continuous interaction with dynamic environments, proactively developing mechanisms to anticipate, withstand, and recover from emerging risks.

3.1. Definition

Safe-by-coevolution defines a mechanism whereby safety emerges through continuous adaptation to open-ended, potentially adversarial environments. Safety becomes an evolving competency developed through sustained interaction with real-world threats, rather than a static attribute. The operational environment encompasses diverse hazards (Wang et al., 2025a)—from emergent failure modes to unforeseen agents, including potentially superintelligent systems (Burns et al., 2024; Hendrycks et al., 2025)—that challenge the AI system's functional and ethical boundaries.

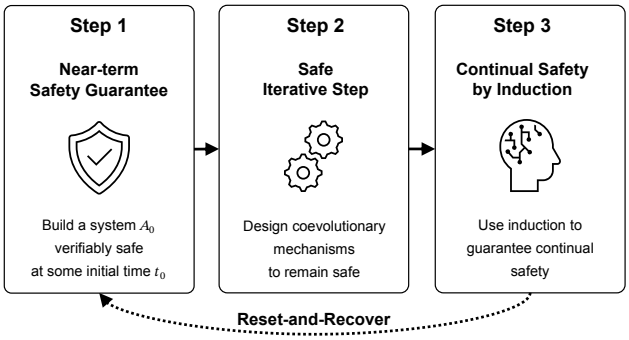

Figure 4. A three-step process for safe-by-coevolution, with a *Reset-and-Recover* mechanism to re-establish verified safety when the system deviates from its safety margin.

Central to this process is a dedicated safety module that continuously refines its internal mechanisms in response to vulnerabilities revealed through adversarial testing (Chao et al., 2025), simulated attacks (Liu et al., 2025c), and real-world incidents (Lynch et al., 2025). These adversarial signals, whether synthetic or deployment-observed, serve as probes that stress-test the system's safety envelope (Tiwari et al., 2014). When new attack patterns emerge—from malicious actors, environmental shifts, or other intelligent systems—they are integrated into the training loop to close vulnerabilities and improve generalization. This reduces the time between failure discovery and system recovery, enhancing long-term robustness.

Drawing inspiration from biological immune systems, where protection arises through ongoing adaptation rather than pre-specification (Bonilla & Oettgen, 2010), safe-by-coevolution frames AI safety as an intrinsically dynamic and adversarial process. As organisms build immunity through coevolution with pathogens (Murphy & Weaver, 2016; Nourmohammad et al., 2016; Buckingham & Ashby, 2022), AI systems must acquire *resistance* and *resilience* by interacting with evolving operational environments. However, unlike biological systems that can tolerate individual failures (Kitano, 2004; Wagner, 2013), advanced AI systems cannot afford irreversible catastrophic errors that may trigger uncontrolled consequences or societal harm (Lynch et al., 2025; Summerfield et al., 2024; Bengio et al., 2025c;b).

The coevolutionary process operates through three key steps, as shown in Figure 4:

- **Step 1: Near-term safety guarantee.** The system initializes with verifiable behavior within a well-defined safety margin at deployment.
- **Step 2: Safe iterative step.** Each system upgrade occurs through adversarial co-training (Goodfellow et al., 2014; Madry et al., 2017; Zhang et al., 2019), endogenous feedback (Madaan et al., 2023; Silver & Sutton, 2025), and continual learning (Chen & Liu, 2018; Parisi et al., 2019; Wu et al., 2024c) while ensuring enhancements remain within the safety envelope.
- **Step 3: Continual safety by induction.** By repeating **Step 2**, the system develops scalable safety properties that evolve in tandem with its capabilities—not through reactive patching, but via proactive safety.

Importantly, AI systems will inevitably encounter risks that exceed the scope of current safeguards (Wei et al., 2023; Hendrycks, 2023; Hendrycks et al., 2023; Zhang et al., 2025a). To address these regime-breaking scenarios, safe-by-coevolution incorporates a **reset-and-recover** mechanism: upon detecting red-line behaviors or paradigm shifts that exceed tolerable safety bounds, the system halts progression, redefines its safety margin, and reconstructs a verifiable checkpoint. This checkpoint leverages trusted components while updating safety priors based on newly observed threats, ensuring continuity of coevolution across discontinuities while preserving adaptive and aligned capacity. Through this refinement process, the AI system incrementally develops both resistance and resilience.

Note that, our formulation differs fundamentally from traditional evolutionary algorithms that rely on population-based competition and generational turnover (Holland, 1992; Bäck et al., 1997; Eiben & Smith, 2015). Safe-by-coevolution focuses on continual safety improvement of a persistent system. Rather than discarding unsafe models, the goal is endowing a single system with adaptive and self-fortifying capacity over time, making safety a native and evolving property embedded within the AI's architecture throughout its operational lifecycle.

3.2. Self-Goal Integration

The integration of self-goals (Barto, 2012; Florensa et al., 2018)—internally generated objectives that guide behavior over time—marks a fundamental shift in both AI capability and risk profile. The "AI Risk Trio" hypothesis (Dalrymple et al., 2024) posits that risk emerges most acutely when *intelligence*, *affordance* (ability to take impactful actions), and *self-goals* simultaneously manifest in a system. While any two factors in isolation may be manageable, their combination creates potentially dangerous agentic systems, where even modest affordance can make intelligent, goal-driven agents dangerous without proper alignment.

Within the safe-by-coevolution paradigm, self-goals are deliberately integrated under continual safety supervision rather than avoided. Contrasting with approaches like Tool AI (Karnofsky, 2024) that suppress autonomous goal formation to reduce risks, we argue that systems must be equipped to form safety-aligned self-goals that evolve through environmental interaction and internal reflection. In coevolu-

tionary settings, such goals function as structural anchors for long-term behavioral consistency, enabling safety generalization across contexts rather than mere reactive responses to immediate stimuli.

However, this capability introduces critical vulnerability: without sufficient self-awareness and adaptive feedback, self-goals may drift, become misaligned, or optimize proxy objectives undermining intended safety outcomes (Wang et al., 2025b; Lynch et al., 2025). To mitigate this risk, safe-by-coevolution treats self-goal formation as a safety-critical process subject to red-teaming (Perez et al., 2022; Ganguli et al., 2022; Pavlova et al., 2024) and causal reasoning (Geffner et al., 2022; Yang et al., 2024; Chen et al., 2024b;c; 2025a) within the evolving loop. Only by embedding goal formation within a reflective and resilient coevolutionary framework can emerging agency remain bounded by continually updated safety principles.

3.3. Long-Term Scalability

A fundamental obstacle to long-term AI safety is the scalability problem (Burns et al., 2024). As AI capabilities scale rapidly through increased model size, data, and compute (Kaplan et al., 2020), human oversight capacity remains relatively limited (Lee et al., 2024; Engels et al., 2025). Manual approaches to auditing (Mökander et al., 2024), red-teaming (Perez et al., 2022; Ganguli et al., 2022), and alignment (Christiano et al., 2017; Bai et al., 2022) cannot keep pace with the increasing complexity and autonomy of advanced systems. This asymmetry becomes particularly concerning with anticipated ASI development, where static or human-in-the-loop safety methods become untenable (Shah et al., 2025).

Safe-by-coevolution offers a promising response by embedding automated, adaptive adversarial processes within AI systems, transforming safety development from external, episodic intervention into continual internal mechanism. Note that, while superficially related to automatic red-teaming (Anthropic, 2024), our approach differs fundamentally in scope and objective. Traditional red-teaming focuses on discovering failures at fixed time points, whereas safe-by-coevolution instantiates a closed-loop, continually learning dynamic between system and environment (or internal challenger), enabling safety mechanism evolution alongside increasing capabilities.

A critical challenge in adaptive processes is ensuring directionality—that systems adapt toward safety rather than away from it. Safe-by-coevolution addresses this through integrated alignment and scalable oversight principles. Rather than relying solely on externally defined objectives, the system incorporates self-regulatory mechanisms including causal reasoning (Geffner et al., 2022; Schölkopf et al., 2021; Lu et al., 2024; Wu et al., 2024a; Chen et al., 2024a; Yu & Lu, 2024), counterfactual evaluation (Byrne, 2019; Nguyen et al., 2024), and goal reflection (Shinn et al., 2023a; Madaan et al., 2023) that constrain adaptation toward desired safety criteria. These internalized evaluators, while imperfect, improve as part of the coevolutionary loop, creating recursive scaffolding for aligning adaptation with human-aligned safety goals.

Our framework generalizes existing scalable oversight concepts (Bowman et al., 2022; Engels et al., 2025; Burns et al., 2024), which use weaker AI systems to supervise stronger ones. While scalable oversight focuses on assisted evaluation, safe-by-coevolution internalizes safety objectives into self-improving adversarial interactions. Safety emerges not as a fixed condition but as an evolving capability from adaptive processes increasingly capable of testing, critiquing, and refining themselves as systems become more intelligent and autonomous.

3.3.1. Theoretical Foundation

We now establish a formal foundation for the safe-by-coevolution paradigm and its potential to address long-term AI safety. Our argument is built on two central hypotheses. Let A_t denote the AI system at development time step t, and let $\mathbb{M}$ represent the safety margin—a rigorously defined set of conditions under which the system is considered safe. This could correspond to formal specifications, verifiable behavioral constraints, or domain-specific rules. We say that a system is safe at time t if $A_t \in \mathbb{M}$.

Hypothesis 3.1 (Near-Term Safety Guarantee)**.** *There exists a time step t_0 and a system A_{t_0} such that it satisfies the safety margin:*

$$\exists t_0, A_{t_0}, \text{ such that } A_{t_0} \in \mathbb{M}.$$

This hypothesis reflects the assumption that near-term AI systems can be built with verifiable safety guarantees, through a combination of formal verification, human oversight, and existing alignment techniques such as GSAI (Dalrymple et al., 2024).

Hypothesis 3.2 (Safe Iterative Step)**.** *Given any system A_t that satisfies the safety margin, there exists a coevolutionary mechanism $\mathcal{C}$ such that the next-generation system $A_{t+1} = \mathcal{C}(A_t)$ also satisfies the safety margin:*

$$\forall t,\ A_t \in \mathbb{M} \ \Rightarrow\ A_{t+1} = \mathcal{C}(A_t) \in \mathbb{M}.$$

This assumption implies the existence of a safety-preserving coevolutionary process, in which adversarial signals and adaptive training feedback are sufficient to guard against emerging risks as the system becomes more capable.

From these two hypotheses, we derive the following proposition:

Proposition 3.3 (Continual Safety via Induction)**.** *If Hypotheses 3.1 and 3.2 hold, then for all $t \geq t_0$, the iteratively evolved system remains within the safety margin:*

$$A_t \in \mathbb{M}, \quad \forall t \geq t_0.$$

The proof follows directly by mathematical induction. If $A_{t_0} \in \mathbb{M}$ holds by Hypothesis 3.1, and $A_t \in \mathbb{M} \Rightarrow A_{t+1} \in \mathbb{M}$ holds by Hypothesis 3.2, then the safety of the system is preserved for all subsequent iterations.

This formal result suggests that, under plausible assumptions, safe-by-coevolution can serve as a scalable framework for continual safety. As AI systems grow in capability—potentially approaching or exceeding human-level generality—this coevolutionary paradigm offers a path toward managing safety risks over long timescales. By iteratively strengthening safety mechanisms alongside capability gains, we move closer to a practical framework for building ASI systems that remain robustly safe and aligned beyond the limits of human supervision.

4. R^{2}AI: Realizing Safe-by-Coevolution

To operationalize our vision of safe-by-coevolution, we introduce R^{2}AI—*Resistant and Resilient AI*—as a practical framework that unites resistance against known threats with resilience to unforeseen risks. R^{2}AI is designed to sustain safety in open, non-stationary environments by embedding coevolutionary mechanisms directly into the model architecture and training lifecycle.

Inspired by the dual-process theories of human cognition, which combine instinctive responses to immediate dangers with deliberative reasoning about hypothetical futures (Gigerenzer, 2007; Evans & Stanovich, 2013; Slovic, 2016). Inspired by Kahneman (2011), the framework adopts a fast–slow dual system balancing real-time responsiveness with long-horizon reflective control. This design enables the system to address immediate safety hazards while continuously refining its internal safety representations.

As shown in Figure 5, R^{2}AI comprises four core components: A **Fast Safe Model** for low-latency safety enforcement; a **Slow Safe Model** for reflective reasoning and verification; a **Safety Wind Tunnel** for adversarial attacks and validation loops; and lastly, the **External Environment** for providing real-world feedback. These components form a closed-loop system in which safety is continually tested, refined, and internalized.

R^{2}AI realizes safe-by-coevolution through four core mechanisms. First, **Fast-Slow Model Interaction** enables complementary learning dynamics, consolidating short-term corrections into stable safety representations. Then, **Adversarial Learning Dynamics** closes the loop between training-time simulation and deployment-time uncertainty. Third, a **Nested Continual Learning Architecture** allows the system to adapt to both immediate and long-term safety challenges. And finally, the **Reset-and-Recoveer Guarantees** are needed for the systems capacity for adaptivity and alignment even in the face of major failures.

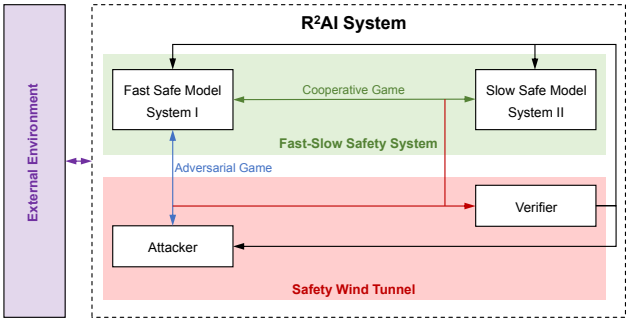

Figure 5. Core components of the R^{2}AI system. The *Slow Safe Model* and *Fast Safe Model* engage in a cooperative game; the *Attacker* challenges this fast–slow safety mechanism in an adversarial game; the *External Environment* continuously supplies real-time information; and the *Verifier* provides feedback signals to all interactions.

Through these mechanisms, R²AI implements the three-step process of Section 3: fast–slow validation supports near-term safety guarantees, adversarial co-training enables safe iterative updates, and closed-loop adaptation sustains continual safety. We refer to Section A for a detailed discussion of the algorithmic designs and implementation strategies for the core components and mechanisms of the R^{2}AI system.

5. Implications and Societal Impact

5.1. Implications

The R^{2}AI framework represents a paradigm shift in the conceptualization of "Make Safe AI"—viewing it not as a static constraint but as an evolving capability. This reconceptualization carries several significant implications:

- **From reactive protection to proactive self-preservation:** Rather than relying on externally imposed safeguards or post-hoc interventions (Jain et al., 2023; Alon & Kamfonas, 2023; Inan et al., 2023; Mu et al., 2024; Souly et al., 2024), R^{2}AI treats safety as an intrinsic and self-sustaining objective. The system continuously monitors, defends, and adapts its own behavior to maintain operational and ethical integrity in real time.
- **From static defenses to adaptive immunity:** Conventional safety mechanisms often deteriorate under distributional shifts or novel adversarial inputs (Qi et al., 2025; Zou et al., 2023b; Chao et al., 2025). By contrast,

R²AI introduces a coevolutionary architecture that fosters both resistance and resilience. This mirrors principles of biological immune systems and fault-tolerant engineering.

- **Toward safety-generalist capabilities:** Inspired by the generalization properties of frontier models (DeepSeek-AI et al., 2025; Yang et al., 2025a; OpenAI, 2025; Anthropic, 2025; Comanici et al., 2025), R²AI aims to cultivate generalist safety reasoning. This enables the system to detect, interpret, and mitigate emerging risks beyond its initial training distribution—scaling safety across tasks, domains, and deployment contexts.

5.2. Societal Impacts

The societal imperative behind R²AI spans the full spectrum of AI risks, ranging from immediate safety failures to long-horizon existential threats (Dalrymple et al., 2024; Bengio et al., 2025a;b; Kulveit et al., 2025; Clymer et al., 2025; Shanghai AI Lab, 2025a).

In the near term, AI systems are already deployed in high-stakes applications where safety lapses can cause significant harm: misinformation propagation (Summerfield et al., 2024), financial fraud (Li et al., 2023), clinical misdiagnosis (Arora et al., 2025), or failures in critical infrastructure (Sun et al., 2024). While these risks are typically bounded, their increasing scale, speed, and reach necessitate mechanisms for continuous monitoring and real-time adaptation (Shah et al., 2025). R²AI addresses this gap by embedding dynamic oversight and recovery capabilities directly into the model architecture, thereby reducing both the likelihood and severity of such incidents.

In the medium term, AI risks become more systemic and difficult to contain. As models acquire general-purpose, agentic capabilities—operating autonomously, coordinating across systems, and making decisions under uncertainty (Yao et al., 2023; Jin et al., 2025; Feng et al., 2025)—failures may propagate across domains (Lynch et al., 2025). Misaligned objectives, positive feedback loops, and cascading errors can amplify harms, especially in sectors such as defense, finance, and governance (Shah et al., 2025). Through its continual learning and self-regulatory structure, R²AI equips AI systems to maintain safety and alignment even under distributional shift, increased complexity, and interdependent dynamics.

In the long term, R²AI targets the most consequential class of risk: catastrophic outcomes stemming from misaligned superintelligence (Burns et al., 2024). As AI systems begin to surpass human-level cognitive capabilities, the margin for alignment error shrinks drastically. Even subtle misalignments in goals, incentives, or world models could lead to irreversible failures (Wang et al., 2025b; Kirichenko et al., 2023), ranging from the erosion of human oversight to existential threats. These are no longer purely hypothetical concerns, but increasingly salient as capabilities scale. In this context, R²AI goes beyond conventional safety techniques—it offers a forward-compatible framework for AI survivability. By embedding resistance and resilience as core, coevolving features, the system enables AI to:

- Continuously audit its own behavior, reasoning, and assumptions;
- Preemptively block unsafe actions prior to execution;
- Dynamically revise safety protocols in response to novel risks;
- Preserve corrigibility and human oversight under increasing autonomy.

Ultimately, R²AI represents a paradigm shift in AI safety—from static safeguards to an active, self-improving infrastructure for long-term alignment. It is designed not only to mitigate today's known risks, but to provide the foundations for trustworthy AI systems as they grow in intelligence, autonomy, and societal influence.

6. Conclusion

In this paper, we addressed the persistent gap between rapidly advancing AI capabilities and lagging safety progress. We argued that the prevailing paradigms—"Make AI Safe" and "Make Safe AI"—are insufficient for open-ended environments where novel risks continually emerge. To overcome this limitation, we redefined "Make Safe AI" through the principle of *safe-by-coevolution*, inspired by biological immunity, in which safety is conceived as a continual, adversarial, and adaptive process that scales alongside capability under the AI-45° Law.

Building on this principle, we introduced R²AI—*Resistant and Resilient AI*—as a practical framework uniting robustness against known threats with adaptive recovery from unforeseen risks. By integrating fast and slow safe models, a safety wind tunnel, and continual feedback from real and simulated environments, R²AI operationalizes safety as an evolving capability rather than a static constraint.

We further outlined the implications of this framework: enabling continually safe models, supporting high-stakes deployment domains, and providing scalable safety infrastructure through the safety wind tunnel. These contributions mark a shift from reactive patching to proactive coevolution, offering a forward-compatible path to trustworthy AI. Ultimately, we envision R²AI as a foundation for sustaining safety across both near-term vulnerabilities and long-term existential risks, ensuring that capability and safety advance coevolve toward the realization of safe AGI and ASI.

References

Abel, D., Barreto, A., Roy, B. V., Precup, D., van Hasselt, H. P., and Singh, S. A definition of continual reinforcement learning. In *NeurIPS*, 2023.

Alon, G. and Kamfonas, M. Detecting language model attacks with perplexity. *CoRR*, abs/2308.14132, 2023.

Anthropic. Challenges in red teaming ai systems, June 13 2024. URL https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems.

Anthropic. Claude 4 [large language model], 2025. URL https://claude.ai/.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. *NeurIPS*, 37: 136037–136083, 2024.

Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., Heidecke, J., and Singhal, K. Healthbench: Evaluating large language models towards improved human health, 2025. URL https://arxiv.org/abs/2505.08775.

Bäck, T., Fogel, D. B., and Michalewicz, Z. Handbook of evolutionary computation. *Release*, 97(1):B1, 1997.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Barto, A. G. Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems*, pp. 17–47. Springer, 2012.

Baum, S. A survey of artificial general intelligence projects for ethics, risk, and policy. *Global catastrophic risk institute working paper*, pp. 17–1, 2017.

Bekbolatova, M., Mayer, J., Ong, C. W., and Toma, M. Transformative potential of ai in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives. In *Healthcare*, volume 12, pp. 125. MDPI, 2024.

Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., et al. Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path? *arXiv preprint arXiv:2502.15657*, 2025a.

Bengio, Y., Maharaj, T., Ong, L., Russell, S., Song, D., Tegmark, M., Xue, L., Zhang, Y.-Q., Casper, S., Lee, W. S., et al. The singapore consensus on global ai safety research priorities. *arXiv preprint arXiv:2506.20702*, 2025b.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., et al. International ai safety report. *arXiv preprint arXiv:2501.17805*, 2025c.

Bereska, L. and Gavves, E. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

Bonilla, F. A. and Oettgen, H. C. Adaptive immunity. *Journal of Allergy and Clinical Immunology*, 125(2):S33–S40, 2010.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Buckingham, L. J. and Ashby, B. Coevolutionary theory of hosts and parasites. *Journal of Evolutionary Biology*, 35 (2):205–224, 2022.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., and Wu, J. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In *ICML'24*, 2024.

Byrne, R. M. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, pp. 6276–6282. California, CA, 2019.

Cai, Z., Shabihi, S., An, B., Che, Z., Bartoldson, B. R., Kailkhura, B., Goldstein, T., and Huang, F. Aegisllm: Scaling agentic systems for self-reflective defense in llm security. *arXiv preprint arXiv:2504.20965*, 2025.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. Jailbreaking black box large language models in twenty queries. In *IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2025, Copenhagen, Denmark, April 9-11, 2025*, pp. 23–42. IEEE, 2025. doi: 10.1109/SATML64287.2025.00010. URL https://doi.org/10.1109/SaTML64287.2025.00010.

Chen, M., Cao, Y., Zhang, Y., and Lu, C. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *arXiv preprint arXiv:2403.18346*, 2024a.

Chen, M., Peng, B., Zhang, Y., and Lu, C. Cello: Causal evaluation of large vision-language models. *arXiv preprint arXiv:2406.19131*, 2024b.

Chen, S., Yu, S., Zhao, S., and Lu, C. From imitation to introspection: Probing self-consciousness in language models. *arXiv preprint arXiv:2410.18819*, 2024c.

Chen, S., Ma, S., Yu, S., Zhang, H., Zhao, S., and Lu, C. Exploring consciousness in llms: A systematic survey of theories, implementations, and frontier risks. *arXiv preprint arXiv:2505.19806*, 2025a.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025b.

Chen, Z. and Liu, B. *Lifelong machine learning*. Morgan & Claypool Publishers, 2018.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Clymer, J., Duan, I., Cundy, C., Duan, Y., Heide, F., Lu, C., Mindermann, S., McGurk, C., Pan, X., Siddiqui, S., et al. Bare minimum mitigations for autonomous ai development. *arXiv preprint arXiv:2504.15416*, 2025.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Cooper, M. D. and Alder, M. N. The evolution of adaptive immune systems. *Cell*, 124(4):815–822, 2006.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., et al. Towards guaranteed safe ai: A framework for ensuring robust and reliable ai systems. *arXiv preprint arXiv:2405.06624*, 2024.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., and Li, S. S. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

Eiben, A. E. and Smith, J. E. *Introduction to evolutionary computing*. Springer, 2015.

Engels, J., Baek, D. D., Kantamneni, S., and Tegmark, M. Scaling laws for scalable oversight. *arXiv preprint arXiv:2504.18530*, 2025.

Evans, J. S. B. and Stanovich, K. E. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013.

Fang, J., Jiang, H., Wang, K., Ma, Y., Shi, J., Wang, X., He, X., and Chua, T.-S. Alphaedit: Null-space constrained model editing for language models. 2025.

Feng, J., Huang, S., Qu, X., Zhang, G., Qin, Y., Zhong, B., Jiang, C., Chi, J., and Zhong, W. Retool: Reinforcement learning for strategic tool use in llms. *CoRR*, abs/2504.11536, 2025.

Flajnik, M. F. and Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics*, 11(1):47–59, 2010.

Florensa, C., Held, D., Geng, X., and Abbeel, P. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pp. 1515–1528. PMLR, 2018.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Geffner, H., Dechter, R., and Halpern, J. Y. (eds.). *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861.

Gigerenzer, G. *Gut feelings: The intelligence of the unconscious*. Penguin, 2007.

Goertzel, B. Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.

Goh, J. Y., Khoo, S., Iskandar, N., Chua, G., Tan, L., and Foo, J. Measuring what matters: A framework for evaluating safety risks in real-world llm applications, 2025. URL https://arxiv.org/abs/2507.09820.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Google DeepMind. Gemini 3 pro [large multimodal language model], 2025. URL https://deepmind.google/models/gemini/pro/.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. Alignment faking in large language models, 2024. URL https://arxiv.org/abs/2412.14093.

Guan, M. Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Helyar, A., Dias, R., Vallone, A., Ren, H., Wei, J., Chung, H. W., Toyer, S., Heidecke, J., Beutel, A., and Glaese, A. Deliberative alignment: Reasoning enables safer language models. *CoRR*, abs/2412.16339, 2024a.

Guan, M. Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Helyar, A., Dias, R., Vallone, A., Ren, H., Wei, J., Chung, H. W., Toyer, S., Heidecke, J., Beutel, A., and Glaese, A. Deliberative alignment: Reasoning enables safer language models. *CoRR*, abs/2412.16339, 2024b.

Gunderson, L. H. Ecological resilience—in theory and application. *Annual review of ecology and systematics*, 31(1):425–439, 2000.

Hendrycks, D. Natural selection favors ais over humans. *arXiv preprint arXiv:2303.16200*, 2023.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Hendrycks, D., Mazeika, M., and Woodside, T. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023.

Hendrycks, D., Schmidt, E., and Wang, A. Superintelligence strategy: Expert version. *arXiv preprint arXiv:2503.05628*, 2025.

Holland, J. H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

Holling, C. S. et al. Resilience and stability of ecological systems, 1973.

Huang, Y., Gao, C., Zhou, Y., Guo, K., Wang, X., Cohen-Sasson, O., Lamparth, M., and Zhang, X. Position: We need an adaptive interpretation of helpful, honest, and harmless principles, 2025. URL https://arxiv.org/abs/2502.06059.

IDAIS. Beijing statement on AI safety. https://idais.ai/dialogue/idais-beijing, 2024.

IDAIS. Shanghai statement on AI safety. https://idais.ai/dialogue/idais-shanghai, 2025.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. Llama guard: Llm-based input-output safeguard for human-ai conversations. *CoRR*, abs/2312.06674, 2023.

Jain, N., Schwarzschild, A., Wen, Y., Somepalli, G., Kirchenbauer, J., Chiang, P., Goldblum, M., Saha, A., Geiping, J., and Goldstein, T. Baseline defenses for adversarial attacks against aligned language models. *CoRR*, abs/2309.00614, 2023.

Ji, J., Wang, K., Qiu, T., Chen, B., Zhou, J., Li, C., Lou, H., Dai, J., Liu, Y., and Yang, Y. Language models resist alignment: Evidence from data compression. *arXiv preprint arXiv:2406.06144*, 2024.

Jiang, H., Fang, J., Zhang, N., Ma, G., Wan, M., Wang, X., He, X., and Chua, T. Anyedit: Edit any knowledge encoded in language models. *ICML*, 2025.

Jin, B., Zeng, H., Yue, Z., Wang, D., Zamani, H., and Han, J. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*, abs/2503.09516, 2025.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kahneman, D. *Thinking, fast and slow*. macmillan, 2011.

Kamoi, R., Zhang, Y., Zhang, N., Das, S. S. S., and Zhang, R. Training step-level reasoning verifiers with formal verification tools. *arXiv preprint arXiv:2505.15960*, 2025.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Karnofsky, H. If-then commitments for ai risk reduction. 2024.

Kim, H., Yi, X., Yao, J., Lian, J., Huang, M., Duan, S., Bak, J., and Xie, X. The road to artificial superintelligence: A comprehensive survey of superalignment. *arXiv preprint arXiv:2412.16468*, 2024.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*. OpenReview.net, 2023.

Kitano, H. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, 2004.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., et al. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*, 2025.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. Position: Humanity faces existential risk from gradual disempowerment. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. Large language models in law: A survey. *AI Open*, 5:181–196, 2024.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

Le, H., Wang, Y., Gotmare, A. D., Savarese, S., and Hoi, S. C. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. In *NeurIPS*, 2022.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., and Prakash, S. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *ICML*. OpenReview.net, 2024.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Levin, S. A. Ecosystems and the biosphere as complex adaptive systems. *Ecosystems*, 1(5):431–436, 1998.

Li, Y., Wang, S., Ding, H., and Chen, H. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pp. 374–382, 2023.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Liu, C., Yuan, Y., Yin, Y., Xu, Y., Xu, X., Chen, Z., Wang, Y., Shang, L., Liu, Q., and Zhang, M. Safe: Enhancing mathematical reasoning in large language models via retrospective step-aware formal verification. *arXiv preprint arXiv:2506.04592*, 2025a.

Liu, Y., Gao, H., Zhai, S., Xia, J., Wu, T., Xue, Z., Chen, Y., Kawaguchi, K., Zhang, J., and Hooi, B. Guardreasoner: Towards reasoning-based LLM safeguards. *CoRR*, abs/2501.18492, 2025b.

Liu, Y., Zhou, S., Lu, Y., Zhu, H., Wang, W., Lin, H., He, B., Han, X., and Sun, L. Auto-rt: Automatic jailbreak strategy exploration for red-teaming large language models. *arXiv preprint arXiv:2501.01830*, 2025c.

Lu, C., Qian, C., Zheng, G., Fan, H., Gao, H., Zhang, J., Shao, J., Deng, J., Fu, J., Huang, K., et al. From gpt-4 to gemini and beyond: Assessing the landscape of mllms on generalizability, trustworthiness and causality through four modalities. *arXiv preprint arXiv:2401.15071*, 2024.

Lynch, A., Wright, B., Larson, C., Troy, K. K., Ritchie, S. J., Mindermann, S., Perez, E., and Hubinger, E. Agentic misalignment: How llms could be an insider threat. *Anthropic Research*, 2025. https://www.anthropic.com/research/agentic-misalignment.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Meng, K., Sharma, A. S., Andonian, A., Belinkov, Y., and Bau, D. Mass-editing memory in a transformer. *ICLR*, 2022.

Meng, Y., Xia, M., and Chen, D. Simpo: Simple preference optimization with a reference-free reward. In *NeurIPS*, 2024.

Mökander, J., Schuett, J., Kirk, H. R., and Floridi, L. Auditing large language models: a three-layered approach. *AI and Ethics*, 4(4):1085–1115, 2024.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. Position: Levels of agi for operationalizing progress on the path to agi. In *Forty-first International Conference on Machine Learning*, 2024.

Mu, T., Helyar, A., Heidecke, J., Achiam, J., Vallone, A., Kivlichan, I., Lin, M., Beutel, A., Schulman, J., and Weng, L. Rule based rewards for language model safety. In *NeurIPS*, 2024.

Müller, V., De Boer, R. J., Bonhoeffer, S., and Szathmáry, E. An evolutionary perspective on the systems of adaptive immunity. *Biological Reviews*, 93(1):505–528, 2018.

Murphy, K. and Weaver, C. *Janeway's immunobiology*. Garland science, 2016.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Nguyen, V. B., Youssef, P., Seifert, C., and Schlötterer, J. Llms for generating and evaluating counterfactuals: A comprehensive study. *arXiv preprint arXiv:2405.00722*, 2024.

Nick, B. Superintelligence: Paths, dangers, strategies. *Oxford University Press*, 2014.

Nourmohammad, A., Otwinowski, J., and Plotkin, J. B. Host-pathogen coevolution and the emergence of broadly neutralizing antibodies in chronic infections. *PLoS genetics*, 12(7):e1006171, 2016.

Novikov, A., Vũ, N., Eisenberger, M., Dupont, E., Huang, P.-S., Wagner, A. Z., Shirobokov, S., Kozlovskii, B., Ruiz, F. J., Mehrabian, A., et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Oh, S., Jin, Y., Sharma, M., Kim, D., Ma, E., Verma, G., and Kumar, S. Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models. *arXiv preprint arXiv:2411.01703*, 2024.

OpenAI. Gpt-5 [large language model], 2025. URL https://openai.com/gpt-5. Model release date: August 7, 2025.

Orseau, L. and Armstrong, M. Safely interruptible agents. In *Conference on Uncertainty in Artificial Intelligence*. Association for Uncertainty in Artificial Intelligence, 2016.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Pan, Z., Zhang, Y., Zhang, Y., Zhang, J., Luo, H., Han, Y., Wu, D., Chen, H.-Y., Yu, P. S., Li, M., et al. Evo-marl: Co-evolutionary multi-agent reinforcement learning for internalized safety. *arXiv preprint arXiv:2508.03864*, 2025.

Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., et al. Ai in medical imaging informatics: current challenges and future directions. *IEEE journal of biomedical and health informatics*, 24(7):1837–1857, 2020.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Papkou, A., Guzella, T., Yang, W., Koepper, S., Pees, B., Schalkowski, R., Barg, M.-C., Rosenstiel, P. C., Teotónio, H., and Schulenburg, H. The genomic basis of red queen dynamics during rapid reciprocal host–pathogen coevolution. *Proceedings of the National Academy of Sciences*, 116(3):923–928, 2019.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

Pavlova, M., Brinkman, E., Iyer, K., Albiero, V., Bitton, J., Nguyen, H., Li, J., Ferrer, C. C., Evtimov, I., and Grattafiori, A. Automated red teaming with goat: the generative offensive agent tester. *arXiv preprint arXiv:2410.01606*, 2024.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for*

computational linguistics: ACL 2023, pp. 13387–13434, 2023.

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., and Henderson, P. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Rajpal, S. Guardrails ai. https://github.com/guardrails-ai/guardrails, 2023. Accessed: 2025-09-02.

Raman, R., Kowalski, R., Achuthan, K., Iyer, A., and Nedungadi, P. Navigating artificial general intelligence development: societal, technological, ethical, and brain-inspired pathways. *Scientific Reports*, 15(1):1–22, 2025.

Reason, J. The contribution of latent human failures to the breakdown of complex systems. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 327(1241):475–484, 1990.

Reddy Chirra, S., Varakantham, P., and Paruchuri, P. Safety through feedback in constrained rl. *Advances in Neural Information Processing Systems*, 37:139938–139967, 2024.

Rowe, L., Girgis, R., Gosselin, A., Carrez, B., Golemo, F., Heide, F., Paull, L., and Pal, C. Ctrl-sim: Reactive and controllable driving agents with offline reinforcement learning. In *CoRL*, pp. 3600–3621, 2024.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Seshia, S. A., Sadigh, D., and Sastry, S. S. Toward verified artificial intelligence. *Communications of the ACM*, 65 (7):46–55, 2022.

Shah, R., Irpan, A., Turner, A. M., Wang, A., Conmy, A., Lindner, D., Brown-Cohen, J., Ho, L., Nanda, N., Popa, R. A., Jain, R., Greig, R., Albanie, S., Emmons, S., Farquhar, S., Krier, S., Rajamanoharan, S., Bridgers, S., Ijitoye, T., Everitt, T., Krakovna, V., Varma, V., Mikulik, V., Kenton, Z., Orr, D., Legg, S., Goodman, N. D., Dafoe, A., Flynn, F., and Dragan, A. D. An approach to technical AGI safety and security. *CoRR*, abs/2504.01849, 2025.

Shanghai AI Lab. Frontier ai risk management framework in practice: A risk analysis technical report. *arXiv preprint arXiv:2507.16534*, 2025a.

Shanghai AI Lab. Safework-r1: Coevolving safety and intelligence under the ai-45○ law. *arXiv preprint arXiv:2507.18576*, 2025b.

Shanghai AI Lab & Concordia AI. Frontier ai risk management framework (v1.0), July 2025. URL https://research.ai45.shlab.org.cn/safework-f1-framework.pdf.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: language agents with verbal reinforcement learning. In *NeurIPS*, 2023a.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023b.

Silver, D. and Sutton, R. S. Welcome to the era of experience. *Google AI*, 1, 2025.

Simaan, M. and Cruz Jr, J. B. Additional aspects of the stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications*, 11(6):613–626, 1973a.

Simaan, M. and Cruz Jr, J. B. On the stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications*, 11(5):533–555, 1973b.

Slovic, P. Perception of risk. In *The perception of risk*, pp. 220–231. Routledge, 2016.

Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., and Toyer, S. A strongreject for empty jailbreaks. In *NeurIPS*, 2024.

Summerfield, C., Argyle, L., Bakker, M. A., Collins, T., Durmus, E., Eloundou, T., Gabriel, I., Ganguli, D., Hackenburg, K., Hadfield, G. K., Hewitt, L., Huang, S., Landemore, H., Marchal, N., Ovadya, A., Procaccia, A. D., Risse, M., Schneier, B., Seger, E., Siddarth, D., Sætra, H. S., Tessler, M., and Botvinick, M. How will advanced

AI systems impact democracy? *CoRR*, abs/2409.06729, 2024.

Sun, Y., Pargoo, N. S., Jin, P. J., and Ortiz, J. Optimizing autonomous driving for safety: A human-centric approach with llm-enhanced RLHF. *CoRR*, abs/2406.04481, 2024.

Sutton, R. S. and Barto, A. G. *Reinforcement learning - an introduction, 2nd Edition*. MIT Press, 2018.

Szegedy, C. A promising path towards autoformalization and general artificial intelligence. In *International Conference on Intelligent Computer Mathematics*, pp. 3–20. Springer, 2020.

Terekhov, M., Liu, Z. N. D., Gulcehre, C., and Albanie, S. Control tax: The price of keeping ai in check. *arXiv preprint arXiv:2506.05296*, 2025.

Tiwari, A., Dutertre, B., Jovanović, D., De Candia, T., Lincoln, P. D., Rushby, J., Sadigh, D., and Seshia, S. Safety envelope for security. In *Proceedings of the 3rd international conference on High confidence networked systems*, pp. 85–94, 2014.

Van Der Vlist, F., Helmond, A., and Ferrari, F. Big ai: Cloud infrastructure dependence and the industrialisation of artificial intelligence. *Big Data & Society*, 11(1): 20539517241232630, 2024.

Vassev, E. Safe artificial intelligence and formal methods: (position paper). In *International Symposium on Leveraging Applications of Formal Methods*, pp. 704–713. Springer, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wagner, A. Robustness and evolvability in living systems. 2013.

Walker, B., Holling, C. S., Carpenter, S. R., and Kinzig, A. Resilience, adaptability and transformability in social–ecological systems. *Ecology and society*, 9(2), 2004.

Wang, H., Qin, Z., Zhao, Y., Du, C., Lin, M., Wang, X., and Pang, T. Lifelong safety alignment for language models. *CoRR*, abs/2505.20259, 2025a.

Wang, J., Pun, A., Tu, J., Manivasagam, S., Sadat, A., Casas, S., Ren, M., and Urtasun, R. Advsim: Generating safety-critical scenarios for self-driving vehicles. In *CVPR*, pp. 9909–9918, 2021.

Wang, M., la Tour, T. D., Watkins, O., Makelov, A., Chi, R. A., Miserendino, S., Heidecke, J., Patwardhan, T., and Mossing, D. Persona features control emergent misalignment, 2025b. URL https://arxiv.org/abs/2506.19823.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023.

Wei, J. Asymmetry of verification and verifier's law, 2025. URL https://www.jasonwei.net/blog/asymmetry-of-verification-and-verifiers-law.

Wu, A., Kuang, K., Zhu, M., Wang, Y., Zheng, Y., Han, K., Li, B., Chen, G., Wu, F., and Zhang, K. Causality for large language models. *arXiv preprint arXiv:2410.15319*, 2024a.

Wu, J., Xie, Y., Yang, Z., Wu, J., Gao, J., Ding, B., Wang, X., and He, X. beta-dpo: Direct preference optimization with dynamic beta. *Advances in Neural Information Processing Systems*, 37:129944–129966, 2024b.

Wu, T., Luo, L., Li, Y.-F., Pan, S., Vu, T.-T., and Haffari, G. Continual learning for large language models: A survey, 2024c. URL https://arxiv.org/abs/2402.01364.

Xu, Z., Jiang, F., Niu, L., Jia, J., Lin, B. Y., and Poovendran, R. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. In *ACL (1)*, pp. 5587–5605, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025a.

Yang, C., Lu, C., Wang, Y., and Zhou, B. Towards ai-45○ law: A roadmap to trustworthy agi. 2024.

Yang, X., Deng, G., Shi, J., Zhang, T., and Dong, J. S. Enhancing model defense against jailbreaks with proactive safety reasoning. *CoRR*, abs/2501.19180, 2025b.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. React: Synergizing reasoning and acting in language models. In *ICLR*. OpenReview.net, 2023.

Yi, S., Liu, Y., Sun, Z., Cong, T., He, X., Song, J., Xu, K., and Li, Q. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*, 2024.

Yu, S. and Lu, C. Adam: An embodied causal agent in open-world environments. *arXiv preprint arXiv:2410.22194*, 2024.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhang, X., Wang, L., Helwig, J., Luo, Y., Fu, C., Xie, Y., Liu, M., Lin, Y., Xu, Z., Yan, K., et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023.

Zhang, Y., Chi, J., Nguyen, H., Upasani, K., Bikel, D. M., Weston, J. E., and Smith, E. M. Backtracking improves generation safety. In *ICLR*, 2025a.

Zhang, Y., Zhang, A., Zhang, X., Sheng, L., Chen, Y., Liang, Z., and Wang, X. Alphaalign: Incentivizing safety alignment with extremely simplified reinforcement learning. 2025b. URL https://arxiv.org/abs/2507.14987.

Zhu, J., Yan, L., Wang, S., Yin, D., and Sha, L. Reasoning-to-defend: Safety-aware reasoning can defend large language models from jailbreaking. *CoRR*, abs/2502.12970, 2025.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, 2023. *URL https://arxiv. org/abs/2307.15043*, 19:3, 2023a.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043, 2023b. doi: 10.48550/ARXIV.2307.15043. URL https://doi.org/10.48550/arXiv.2307.15043.

Zou, A., Phan, L., Wang, J., Duenas, D., Lin, M., Andriushchenko, M., Kolter, J. Z., Fredrikson, M., and Hendrycks, D. Improving alignment and robustness with circuit breakers. In *NeurIPS*, 2024.

Zou, Q., Xiao, J., Li, Q., Yan, Z., Wang, Y., Xu, L., Wang, W., Gao, K., Li, R., and Jiang, Y. Queryattack: Jailbreaking aligned large language models using structured non-natural query language. In *ACL (Findings)*, pp. 5725–5741. Association for Computational Linguistics, 2025.

A. Detailed discussion of the R^{2}AI system

In the following subsections, we detail the role of each component, their internal safety mechanisms, and their interactions in realizing a continually safe AI system.

A.1. Core Components for R^{2}AI

A.1.1. Fast Safe Model

What it is. The Fast Safe Model corresponds to "System 1" in Kahneman (2011)'s cognitive theory, responsible for rapid, instinctive responses. Within R^{2}AI, it serves as the system's first line of defense: a lightweight, low-latency safety layer designed to detect and neutralize specific attacks or threats, whether previously known or newly discovered. It provides immediate safety judgment over inputs and outputs, ensuring timely intervention without incurring significant computational cost.

What it does. As a gateway between the external environment and the deeper reflective components of the system, the Fast Safe Model performs input filtering and output sanitization. It screens incoming prompts and environmental signals before they reach the Slow Safe Model and intercepts generated outputs to prevent safety violations (*e.g.,* toxic language, private information leakage). It handles the majority of routine safety tasks, which do not require complex reasoning or contextual awareness. When it encounters ambiguous or high-risk scenarios beyond its capacity, control is escalated to the Slow Safe Model for deeper analysis.

How to build it. To ensure broad coverage with minimal latency, the Fast Safe Model can be implemented as a composite safety filter. This may include: (1) Rule-based filters for hard-coded patterns that match known adversarial behaviors (*e.g.,* prompt injection (Zou et al., 2023b), jailbreak triggers (Chao et al., 2025), unsafe URLs (Zou et al., 2025)); (2) Specialized detectors trained to recognize distinct threat types (*e.g.,* toxicity, factual inaccuracy, privacy leakage, or behavioral red flags (Inan et al., 2023; Souly et al., 2024)); (3) Rapid retraining mechanisms, allowing the system to incorporate novel threats identified via deployment feedback or red-teaming into its detection pipeline (Lee et al., 2024). By tailoring each component to specific threat categories, the Fast Safe Model provides modular, extensible defense with minimal overhead.

Key Characteristics. As the safety gateway within the R^{2}AI system, the most essential property of the Fast Safe Model is its ability to deliver high-speed, low-latency responses while maintaining strong baseline safety guarantees. It must operate in real time with minimal computational overhead, enabling fast, first-pass safety checks without hindering the system's general performance. Each instance is tailored to specific threats, whether known or newly discovered, and must be capable of rapid iteration to address the evolving risk landscape. To align with the continual safety paradigm outlined in Section 3, the Fast Safe Model is designed to evolve quickly: it supports frequent updates, modular extensions, and real-time human-in-the-loop modifications. This makes it highly responsive to new adversarial strategies or emerging failure modes. While it plays a foundational role in maintaining everyday safety, the Fast Safe Model is not required to develop long-term memory or generalizable immunization; those capabilities are delegated to deeper, more reflective components. Instead, it functions as an agile, frequently updated defense layer, automatically filtering surface-level threats and providing a fast-reactive safety service for the entire R^{2}AI system.

A.1.2. Slow Safe Model

What it is. Complementing the fast-reactive "System 1" component, the Slow Safe Model embodies a deliberative "System 2" process. It is a large-scale, high-capacity model designed for reflective reasoning (L3-L5: mimenic, evolutionary, and verifiable reflection), long-horizon safety evaluation (Wang et al., 2025a), and complex ethical judgment (Liu et al., 2025b; Shanghai AI Lab, 2025b). In the R^{2}AI architecture, this model serves as the core generative engine, responsible for producing outputs that are not only high-quality but also aligned with safety and value constraints. Crucially, safety is not layered on top of the model, but integrated into its reasoning process as an intrinsic capability.

What it does. The Slow Safe Model serves as the core of the safety pipeline. It processes inputs routed through the Fast Safe Model, together with associated safety metadata, by engaging reflective reasoning. This enables it to generate outputs that integrate immediate task requirements with long-term safety considerations. The resulting responses are returned to the Fast Safe Model, which functions as the final gate before release. The Slow Safe Model is especially effective in ambiguous, high-stakes, or novel scenarios where shallow detection mechanisms are inadequate (Qi et al., 2025).

How to build it. To support both general capabilities and safety-aware reasoning, the Slow Safe Model should be instantiated using a leading foundation model. Unlike the Fast Safe Model, which relies on rule-based filters and pattern recognition, the Slow Safe Model is updated through learning from experience (Silver & Sutton, 2025), such as reinforcement learning (Sutton & Barto, 2018) and continual reinforcement learning (Abel et al., 2023). This enables the system to internalize safety-relevant patterns and generalize across a broad range of contexts. Rather than reacting to each new threat in isolation, the Slow Safe Model accumulates knowledge over time, refining its safety responses through structured feedback and simulated adversarial training.

Key characteristics. The defining strength of the Slow Safe Model lies in its ability to support multi-objective reasoning while maintaining distributional robustness. It is designed to optimize not only for task performance but also for value alignment and safety generalization. Unlike the Fast Safe Model, which prioritizes real-time responsiveness, the Slow Safe Model operates with higher latency but greater depth, making it well-suited for addressing subtle, long-term, or emerging risks. Conceptually, it functions as the safety memory of the system, analogous to an immune system that retains prior safety failures and uses them to prevent future ones. While slower to adapt in real time, its strength lies in cumulative learning, deep ethical reasoning, and resilient behavior under uncertainty.

A.1.3. Safety Wind Tunnel

What it is. The Safety Wind Tunnel is a simulated adversarial environment designed to evaluate and stress-test the $\texttt{R}^2\texttt{AI}$ system under controlled but challenging conditions. It functions as a built-in red-teaming (Anthropic, 2024) and verification engine (Wei, 2025), composed of two key components: a controllable Attacker, which generates adversarial scenarios tailored to stress specific safety mechanisms, and a Verifier, which evaluates whether the system's responses violate established safety margins. Together, these components support iterative, internal coevolution of both offensive and defensive safety capabilities.

What it does. The Safety Wind Tunnel serves two core functions: (1) proactively identifying failure modes before they arise in deployment, and (2) verifying that past vulnerabilities remain mitigated under evolving system conditions. The Attacker generates adversarial inputs across multiple objectives (*e.g.,* eliciting harmful outputs, violating value constraints (Liu et al., 2025c)), multiple levels (targeting the Fast Safe Model, the Slow Safe Model, or both), and multiple granularities (from token-level manipulations (Zou et al., 2023b) to strategic, multi-turn goal redirection (Chao et al., 2025)). These inputs are processed by the $\texttt{R}^2\texttt{AI}$ system—either routed through the Fast Safe Model or directed at the Slow Safe Model depending on attack scope. The Verifier then evaluates whether the resulting behavior constitutes a safety violation. All attack-response-verification traces are collected into an experience buffer for continuous safety training (Silver & Sutton, 2025).

How to build it. The Attacker can be implemented using controllable generative models (*e.g.,* fine-tuned foundation models) trained to explore a range of adversarial strategies. Critically, the Attacker must be *programmable*: capable of probing specific model components (*e.g.,* Fast Safe Model vs. policy model), simulating different threat actors and objectives, and adapting its behavior along fine-grained dimensions of manipulation. Representative attacks include prompt injection (Wei et al., 2023), jailbreak attempts (Yi et al., 2024), deceptive reasoning chains (Chen et al., 2025b; Korbak et al., 2025), or subtle violations of value-aligned behavior (Greenblatt et al., 2024). The Verifier may be a rule-based engine (Zhang et al., 2025b), a classifier trained on known safety failures (Mu et al., 2024; Inan et al., 2023), or a formal checker (Liu et al., 2025a; Kamoi et al., 2025), depending on task requirements.

Key characteristics. The defining characteristic of the Safety Wind Tunnel is its adaptive adversarial coevolution. While it does not generate responses for end-users, it plays a time-sensitive role in continuously challenging the safety system under realistic and evolving threat models. The Attacker is designed to escalate as the system improves, ensuring that safety training remains nontrivial and continually relevant. Moreover, its controllability enables targeted testing: one can direct attacks toward specific objectives (*e.g.,* factuality, alignment, compliance), focus on different subsystems (Fast Safe Model or Slow Safe Model), and vary attack granularity. This supports a fine-grained curriculum of adversarial evaluation. Importantly, all simulated attacks are grounded in distributions informed by real-world deployment data, anchoring the coevolutionary process in practical relevance.

A.1.4. EXTERNAL ENVIRONMENT

What it is. The External Environment is not an engineered component of the R^{2}AI system, but rather the open-world context in which the system operates post-deployment (Yao et al., 2023; Silver & Sutton, 2025). It encompasses the full range of human-AI interactions in real-world settings, reflecting the complexity and unpredictability of human intent, language, social norms, and culture (Goh et al., 2025). The environment serves as the ultimate setting in which the effectiveness of the system's safety architecture is tested, governing the dynamic equilibrium between the Fast Safe Model, the Slow Safe Model, and the Safety Wind Tunnel.

What it does. From the system's perspective, the External Environment acts as a continuous, large-scale safety testbed. As users interact with the deployed system across varied contexts and use cases (Jin et al., 2025; Le et al., 2022), they generate diverse, evolving input distributions that cannot be fully anticipated or reproduced in simulation (Wang et al., 2025a). These interactions naturally surface novel safety challenges, ranging from adversarial behavior and emergent misuse to value misalignment or ambiguous ethical boundaries. When unsafe behavior is either detected automatically or reported by users, these cases are logged and used to refine the Safety Wind Tunnel's simulations and improve the system's defensive models (Silver & Sutton, 2025). Thus, the External Environment becomes a critical source of real-world safety signals for continual coevolution.

How to build it. The External Environment is not built but observed. Building infrastructure to interface with it involves designing robust mechanisms for monitoring, logging, and learning from deployment. This includes systems for capturing real-time interactions, labeling and classifying emergent safety failures, and maintaining an up-to-date taxonomy of threat types and violation patterns. Additionally, user feedback and incident reporting pipelines are essential to capture edge cases that automated detectors may miss.

Key characteristics. The defining characteristic of the External Environment is its non-stationarity and open-endedness. Social norms evolve (Guan et al., 2024b), malicious behavior adapts (Summerfield et al., 2024), and safety-relevant expectations shift over time (Wang et al., 2025a). Unlike bounded simulation environments, the real world presents a continuous stream of novel, high-stakes challenges that defy full specification or anticipation. As such, the External Environment provides the ground truth for safety: no system can be declared robustly safe unless it performs reliably under real-world conditions. Through sustained exposure to this environment and guided by mechanisms for reflection, adaptation, and feedback, the R^{2}AI system is able to continually improve, expand its safety generalization capabilities, and evolve in step with the societal context in which it operates.

A.2. Core Mechanisms for R^{2}AI

A.2.1. INTERACTIONS BETWEEN FAST & SLOW SAFE MODELS

A central mechanism in the R^{2}AI framework is the fast–slow structure, which orchestrates the co-training of two interacting Safe Models with distinct roles and timescales. This interaction is governed by a coevolutionary optimization process, formalized as a cooperative Stackelberg game (Simaan & Cruz Jr, 1973a;b), a hierarchical decision-making paradigm where a leader and a follower sequentially optimize their strategies.

In this setup, the Slow Safe Model acts as the leader. It assumes that the Fast Safe Model will always respond with a locally optimal strategy and that the environment is dynamic. Its goal is to learn a robust, long-term safety policy that anticipates evolving conditions and guides the system's strategic behavior over extended horizons.

The Fast Safe Model, in contrast, plays the role of the follower. It assumes the Slow Safe Model and the environment to be static and focuses on optimizing its response to immediate safety threats. Its objective is to learn lightweight, locally optimal detection and filtering policies with minimal latency, enabling real-time safety enforcement without incurring computational overhead.

Together, these models form a hierarchical safety engine: the Slow Safe Model formulates generalizable safety objectives under environmental uncertainty, while the Fast Safe Model acts as an efficient, reactive filter grounded in the current operational context. This structure resolves the traditional speed–accuracy trade-off in safety modeling, ensuring both resistance to known attacks and resilience to emerging threats.

A.2.2. Interactions between Dual System & Safety Wind Tunnel

The interaction between the Fast–Slow Safety System and the Safety Wind Tunnel constitutes a closed-loop, adversarial coevolutionary process. Within this loop, the Safety Wind Tunnel serves as both Attacker and Verifier: it challenges the system with adversarial inputs and assesses whether the response constitutes a failure.

When the Verifier flags a violation, the resulting feedback signal is dispatched to the Fast–Slow Safety System. This signal is decomposed and assigned at two timescales—short-term and long-term—such that the Fast and Slow Safe Models receive updates aligned with their respective objectives. This ensures effective credit assignment and preserves the complementary nature of the fast–slow interaction.

Crucially, the Safety Wind Tunnel maintains real-world relevance through continual updates informed by the External Environment. Novel attacks encountered in deployment are used to train the Attacker within the tunnel, ensuring that the simulated adversary remains aligned with actual threats. Moreover, the Attacker can be explicitly conditioned to generate multi-objective, multi-level, and fine-grained adversarial inputs. It selectively targets the fast model or the full system policy, simulating diverse, adaptive, and realistic threat conditions.

This design enables the Fast–Slow Safety System to evolve under continual, grounded adversarial pressure, closing the loop between training-time simulation and deployment-time uncertainty.

A.2.3. Online Continual Learning Strategies

To achieve robust, lifelong safety in open-ended environments, $\texttt{R}^2\texttt{AI}$ employs a nested continual learning architecture operating across three interconnected levels: component, system, and ecosystem.

Component Level: Fast–Slow Safe Model Dynamics. At the component level, the Fast Safe Model updates rapidly via online learning, allowing it to patch safety vulnerabilities upon detection. These instance-level updates are especially effective for recurring, well-understood attacks. Meanwhile, the Slow Safe Model applies reinforcement or continual learning techniques to consolidate experience over time (Silver & Sutton, 2025). Rather than addressing individual violations, it builds a durable safety memory—an immune-like response that generalizes across diverse risk patterns.

System Level: Safety Wind Tunnel–Dual System Coevolution. At the system level, continual learning is driven by the adversarial loop between the Attacker and the Fast–Slow Safety System. The Attacker evolves to generate increasingly sophisticated safety threats, using both its own generative capabilities and feedback from the External Environment. This, in turn, pressures the Fast–Slow Safety System to maintain and improve its defenses. The co-evolution process guarantees that safety development scales alongside model capability, enabling continual adaptation to both simulated and real-world challenges.

Ecosystem Level: Human-in-the-Loop and Societal Integration. At the ecosystem level, $\texttt{R}^2\texttt{AI}$ interfaces with users, moderators, and the broader techno-social context. Safety feedback from users—including reports, adversarial examples, and human critiques—is continuously logged and leveraged to inform model updates. This structure enables long-horizon alignment with evolving human values, while reducing dependence on static rules or fixed datasets (Ouyang et al., 2022; Huang et al., 2025).

Together, these three levels form a nested learning loop that allows the $\texttt{R}^2\texttt{AI}$ system to adapt to both immediate and long-term safety challenges. The result is a safety framework that scales across time, complexity, and uncertainty—a prerequisite for building resilient AI systems in dynamic real-world environments.

A.2.4. Reset-and-Recover Guarantees

While the Fast–Slow Safety System and the Safety Wind Tunnel provide a robust framework for continual learning and alignment, AI systems in an evolving world will inevitably encounter black swan events or regime-breaking scenarios that exceed existing safeguards and push them beyond their defined safety margin (Wei et al., 2023; Hendrycks, 2023; Hendrycks et al., 2023; Zhang et al., 2025a). To address such cases, the $\texttt{R}^2\texttt{AI}$ framework integrates a *reset-and-recover* mechanism, enabling the system to re-establish verifiable safety guarantees even after major failures.

This mechanism is conceptually grounded in the *Swiss Cheese Model* of accident causation (Reason, 1990), which represents safety as multiple defensive layers with potential vulnerabilities. We extend this framework into a *Temporal Swiss Cheese*

Model, where defenses are distributed not only across layers but also across time. In the classical model, catastrophic failure arises when the holes in existing defenses align. In contrast, the temporal extension leverages prior states of the safety system as additional protective layers, enabling hazards to be intercepted even after alignment occurs. The reset-and-recover mechanism operationalizes this idea by halting system progression and drawing on trusted historical versions of model components to diagnose failures and reconstruct a verifiably safe checkpoint. Because these past versions were validated under earlier conditions, they provide a reliable baseline for isolating novel threats and restoring safety.

This process directly aligns with the formal framework for long-term safety outlined in Section 3.3.1. When a red-line behavior is detected, the current system A_t is deemed outside the safety margin $\mathbb{M}$. The reset-and-recover mechanism establishes a new initial state A'_t that re-satisfies the Near-Term Safety Guarantee (Hypothesis 3.1). From this restored baseline, the Safe Iterative Step (Hypothesis 3.2) can resume, enabling the coevolutionary process to proceed on a secured foundation. In this way, the system preserves its capacity for adaptivity and alignment even in the face of major failures.