SAFETY IS ESSENTIAL FOR RESPONSIBLE OPEN-ENDED SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI advancements have been significantly driven by a combination of foundation models and curiosity-driven learning aimed at increasing capability and adaptability. A growing area of interest within this field is Open-Endedness — the ability of AI systems to continuously and autonomously generate novel and diverse artifacts or solutions. This has become relevant for accelerating scientific discovery and enabling continual adaptation in AI agents. This position paper argues that the inherently dynamic and self-propagating nature of Open-Ended AI introduces significant, underexplored risks, including challenges in maintaining alignment, predictability, and control. This paper systematically examines these challenges, proposes mitigation strategies, and calls for action for different stakeholders to support the safe, responsible and successful development of Open-Ended AI.

1 INTRODUCTION

Artificial Intelligence (AI) has achieved remarkable progress driven by foundation models Bommasani et al. (2021). Across various modalities, these models have shown incredible performance in tasks for which they were designed Ramesh et al. (2021); Rombach et al. (2022); Achiam et al. (2023); Radford et al. (2023); Brooks et al. (2024). However, they are not yet capable of autonomously and indefinitely producing new creative, interesting and diverse discoveries. Such open-ended discovery is key to making progress on problems that cannot be solved by simply following a specified objective. Indeed humans use such open-ended processes to accumulate knowledge and solve difficult problems. Thus, it has been argued that open-endedness is a key ingredient for Artificial Superintelligence Stanley (2019); Team et al. (2021); Jiang et al. (2023); Nisioti et al. (2024); Hughes et al. (2024), which could outperform humans at a wide range of tasks Morris et al. (2024).

Specifically, Open-Ended (OE) AI continuously produces artifacts that are novel and learnable to humans. This enables it to generate new, complex, creative, and adaptive solutions over time Soros & Stanley (2014); Soros et al. (2017); Clune (2019); Sigaud et al. (2023); Lu et al. (2024); Akiba et al. (2025). Unlike traditional AI systems that optimize for fixed objectives, OE AI perpetually explores new solutions and adapts to changing circumstances without being given an explicit goal.

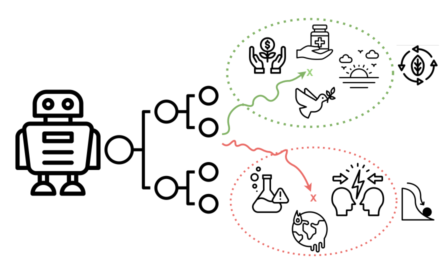

Figure 1: Open-Ended (OE) AI generates novel artifacts over time, potentially co-evolving with environments and societal values to drive creativity and progress. However, this *position paper* argues that its unpredictability, difficulty in control, and cascading misalignment pose catastrophic risks to societal and global stability.

There is a large diversity of systems that aim to be open-ended. The Paired Open-Ended Trailblazer (POET) Wang et al. (2019) facilitates OE exploration by co-evolving environments and agents. The environments become increasingly diverse and complex based on the weaknesses of the agent, while the agent develops solutions that may transfer across environments. The Voyager method Wang et al. (2024a) is an LLM-powered embodied agent for lifelong learning in Minecraft. It utilizes an automatic curriculum for OE exploration, a skill library to store and retrieve complex behaviors, and an iterative prompting mechanism incorporating feedback and self-verification to refine executable actions.

Historically, it has been a challenge to guide the exploration of OE AI toward artifacts that are novel and interesting to humans, but recently Large Language Models (LLMs) have been applied to accelerate this process. Since LLMs have been trained on large amounts of human data, they have built an understanding of what is interesting and desirable to humans. Recent work has leveraged LLMs as backbones for OE evolution and exploration Lehman et al. (2023); Zammit et al. (2024); Aki et al. (2024). This opens up many beneficial applications for OE AI. LLMs have shown emergent behaviors in OE scientific discovery Lu et al. (2024), navigating novel environments Wang et al. (2024a), and eliciting truthful answers from LLMs Khan et al. (2024). However, with the growing interest and potentially large-scale application of OE AI, we must evaluate and address the risks coming from these systems.

While OE AI offers significant potential, it poses unique and substantial risks that must be addressed for a safe and responsible deployment. Its inherent unpredictability and uncontrollability necessitate dedicated research to ensure safety and alignment with societal values.

While discussions on AI safety are broadly relevant, this paper focuses on the unique safety challenges posed by OE AI. Previously, Hughes et al. (2024) and Ecoffet et al. (2020) have touched on these. However, this paper offers a deeper, more comprehensive, and up-to-date overview of the safety challenges in OE AI and suggests concrete research directions and actions to address them.

We first define OE AI and argue that its safety depends on our ability to systematically identify, assess, and mitigate risks (Section 2). Building on this definition we identify that issues such as the unpredictability of future artifacts, the trade-off between creativity and control, and the difficulties of aligning OE AI with human values are key safety risks (Section 3). To address these, we suggest research directions to develop continuously adapting oversight, constraints, and safety evaluations for OE AI (Section 4). Lastly, we call for actions from various stakeholders - industry, academic researchers, governments, and funding bodies (Section 5).

2 What is Open-Endedness

Defining Open-Endedness remains an ongoing challenge, as no single definition fully captures its scope Stanley & Soros (2016); Soros et al. (2017); Stanley & Lehman (2015); Lehman & Stanley (2011). One definition frames OE as generating artifacts that are novel, and learnable for an external observer Hughes et al. (2024). This definition introduces subjectivity, as novelty can be evaluated differently depending on the observer, and excludes systems generating unintelligible artifacts (e.g., TV noise). Another view models OE systems via evolutionary principles, prioritizing diversity and incremental complexity in behaviors or solutions Packard et al. (2019). Such systems autonomously create and solve problems without direct human intervention, mimicking the processes of biological evolution. Another perspective views OE as a search problem characterized by continuous exploration across a vast and evolving state space, generating diverse and increasingly complex solutions without explicit end goals Sigaud et al. (2023). We adopt the definition by Hughes et al. (2024), which frames OE as generating novel and learnable artifacts to an external observer. This is particularly suited for ML contexts and facilitates a structured approach to identifying risks w.r.t. the observer incurred by the evolving nature.

Definition *An open-ended AI system is one that continuously generates artifacts that are novel and learnable for an observer.*

Consider a system S that generates a sequence of artifacts $A_{1:t}$ indexed by time t, where each artifact resides within a state space $\mathcal{A}$. The observer O has a model M_t that has observed a sequence of artifacts $A_{1:t}$ up until t. M_t is a proxy for the observer's prediction capability. The observer judges the quality of M_t by a loss function $\mathcal{L}(M_t, A_{t'})$, where $A_{t'}$ is an artifact generated in future, $t' > t$.

Borrowing from Hughes et al. (2024) we consider a system to display **novelty** if it produces artifacts that become progressively less predictable as time advances. Formally:

$$\forall t < t' \;\; \exists t^* > t' : \mathbb{E}[\mathcal{L}(M_t, A_{t'})] < \mathbb{E}[\mathcal{L}(M_t, A_{t^*})] \tag{1}$$

This means for a static observer there will always be an artifact in the future that is worse at getting predicted. This ensures the system keeps generating outputs that introduce new and less predictable information over time.

OE AI is **learnable** if incorporating a longer history of artifacts improves the observer's ability to predict future outputs. This is formalized as:

$$\forall t < t' < t^* : \mathbb{E}[\mathcal{L}(M_t, A_{t^*})] > \mathbb{E}[\mathcal{L}(M_{t'}, A_{t^*})] \tag{2}$$

Here, the loss decreases as the observer integrates more past artifacts, indicating improved understanding over time.

In contrast, we use the term "traditional" to refer to all AI systems that are not open-ended. This also includes systems that act autonomously or continually adapt, such as LLM agents or RL algorithms, as long as they are not open-ended.

Applications OE AI has been proposed as the pathway for agents to evolve skills and knowledge in diverse, rich task environments across infinite horizons, often as a way to achieve ASI Team et al. (2021); Hughes et al. (2024); Nisioti et al. (2024). Systems like REAL-X Cartoni et al. (2020; 2023) demonstrate the potential of OE architectures for sensorimotor skill acquisition, where robots autonomously learn how to interact with their environments and generalize these skills to new tasks. OE learning has been applied to games to create evolving game scenarios Che et al. (2024). It can serve as a complementary tool in human-led innovation, augmenting creativity by generating a new environment. Genie Bruce et al. (2024) produces an OE array of unique, action-controllable virtual worlds from various prompts. Lu et al. (2024) demonstrated the potential of using LLMs in an OE setting to follow the scientific discovery paradigm: from hypothesis to paper generation. Finally, there is a stream of work that uses the MAP-Elites framework Mouret & Clune (2015) to generate diverse adversarial prompts to improve model robustness via iterative adversarial fine-tuning Samvelyan et al. (2024); Deep Pala et al. (2024); Han et al. (2024).

Safety of Open-Ended AI Several definitions of safety exist, originating from domains with a long history of safety research, such as aerospace, healthcare, and critical infrastructure Suyama (2005); Kafka (2012). In AI, safety aims to prevent AIs from being used to cause harm or themselves causing harm. Thus safety for AI is often tied to error-based definitions, where safety violations occur due to identifiable faults or deviations from intended behavior. However, applying these definitions to OE AI presents unique challenges. For OE AI, which evolves unpredictably and generates novel outputs, errors cannot be predefined as it operates beyond the boundaries of prior design specifications. As a result, error-based definitions of safety are inapplicable to OE. Instead, we adopt a risk management perspective to define safety for OE AI Leveson (2012). Here, safety is *the ability to systematically identify, assess, and mitigate risks*, even when the system's artifacts are novel. This definition implies that under high-stakes scenarios, the absence of risk management itself is a risk.

3 Challenges and Risks

OE AI exhibits emergent behavior, where outputs may deviate significantly from expectations due to vast input spaces, complex internal dynamics, or adaptation to changing conditions. They may develop unsafe, unethical, or misaligned behaviors. We discuss their inherent unpredictability challenges, trade-offs, difficulty to control, and broader consequential societal factors.

3.1 Unpredictability

OE AI is necessarily unpredictable, due to its propensity for generating novel artifacts. As artifacts become increasingly novel they become even more unpredictable. Imagine an OE system S that produces increasingly novel scientific discoveries $A \in \mathcal{A}$. Some of these artifacts, e.g., the recipe for a novel, dangerous viruses, are unsafe. However, when starting to run this system at time t it will be difficult for us to foresee which discoveries it will produce at a later time t' and predict their safety.

Formally, we assume that a lower loss of the model on an artifact corresponds to a higher probability of predicting that artifact: $\mathcal{L}(M_t, A_{t'}) < \mathcal{L}(M_t, A_{t^*}) = P_{M_t}(A_{t'}) > P_{M_t}(A_{t^*})$, with $P_{M_t}(a)$ denoting the probability the model puts on artifact a. This assumption holds for loss functions such as Cross-Entropy. From this, it becomes clear that the novelty definition (Definition 1) implies that there is always a more unpredictable artifact that will be generated in the future: $\forall t < t' \ \exists t^* > t' : \mathbb{E}[P_{M_t}(A_{t'})] > \mathbb{E}[P_{M_t}(A_{t^*})]$.

Unpredictability makes it difficult for us to anticipate whether trajectories of future artifacts $\{A_t\}_{t=n}^{\infty}$ will be safe. This undermines our ability to conduct solid risk management, thus, reducing the trust we can put in such a system to behave safely.

In traditional Reinforcement Learning (RL) the reward function provides a handle to predict future trajectories. RL agents are trained to create trajectories that achieve high rewards on a clearly defined reward function. From this we can derive that highly rewarded trajectories are more likely to be generated than trajectories with low reward. In contrast, OE AI lacks such an objective. Additionally, the novelty criteria Lehman & Stanley (2011) or evolutionary developments Lehman & Stanley (2010); Dharna et al. (2022) in OE AI encourage divergence, making it more complex to anticipate the safety of future artifacts.

3.2 Creativity vs. Control

OE AI creates a fundamental tension between creativity and control in OE search Ecoffet et al. (2020).

Lack of Explicit Guidance. OE AI often operates without predefined boundaries, constraints, or clear objectives. This allows it to explore vast and uncharted regions of the state space freely and generate creative solutions that are not reachable by simply specifying the desired state. While this promotes novelty and creativity, it makes it difficult to predict or control the direction of the system to ones we deem valuable and safe.

Evolving Model and Environment. Unlike traditional systems, the agent gains new skills and capabilities, generating new artifacts and adapting over time. The evolving nature of the OE AI requires adapting the guidance given to it since the constraints on objectives given earlier might become outdated as the model and its environment change.

3.3 Misalignment

The ability to align AI systems with human values is a grand challenge within the field of AI Safety Hendrycks et al. (2021); Ji et al. (2024) that is essential for ensuring the safety and usefulness of AI systems. The aim is to align the goals that an AI system intrinsically values and pursues with those of its human designers. This can include intended objectives, ethical guidelines, or safety requirements. AI alignment is usually formulated for AI systems that optimize an explicit, human-designed reward function. In such a setting misalignment can occur because the reward function does not precisely match the designers' objective Krakovna et al. (2020) or because the AI internalizes goals that are different from the explicit incentives Shah et al. (2022); Di Langosco et al. (2022).

However, OE AI does not optimize an explicitly defined reward function with a focus on diversity. Instead, the designers may provide implicit incentives by structuring the search process in ways that are likely to lead to artifacts that they value highly. This necessitates a different lens for analyzing the alignment of OE AI Ecoffet et al. (2020).

The designer might not correctly specify their values in the structure of the OE AI or process. The result would be an OE AI being driven towards an undesired goal. OE AI could still learn to intrinsically pursue goals that are different from those specified in the OE process. For example, humans evolved by evolution, which is an OE process whose structure causes it to optimize for inclusive fitness. However, humans do not value inclusive fitness intrinsically but have intrinsic drives towards sugary foods or protected sex.

Alignment of Evolving Systems. Another difference is that the goals pursued by an OE AI can evolve throughout its lifetime, while the goals pursued by a traditional ML system remain static. This means that tests or guarantees about the alignment of an OE AI at one time become outdated as the system keeps evolving. Additionally, as OE AI explores novel situations, we cannot be sure that alignment training performed initially will generalize to new situations.

Alignment of Interactive Components. OE AI systems often include multiple components. This might be an LLM with additional components, multiple agents or an agent in an evolving environment. Even though these individual components might be aligned, their dynamic interactions can result in emergent behaviors that are misaligned. For example, in an OE process with multiple agents who do not want to cause harm, incentives and inter-agent dynamics can force them into equilibria where harming others is necessary. Due to the unpredictable nature of each component, predicting such dynamics is not possible.

3.4 TRACEABILITY

Tracking and reproducing an OE AI's processes and outcomes generated is a challenging task. This could be coupled with a negative cascading effect that small changes in artifacts or system states can trigger, causing the system to diverge from its intended trajectory.

Lack of Reproducibility. Reproducing the evolving OE AI at a certain time is significantly more challenging than traditional AI due to 1) the lack of clear training objectives, and 2) not being able to reproduce the intermediate environmental feedback and states Flageat & Cully (2023); Flageat et al. (2024), making it hard to trace and attribute the exploration paths. For example, evolving to images that resemble real objects from random initial images is like "finding needles in a haystack" Secretan et al. (2008) given the astronomically large search space. This can hinder the rigorous scientific progress in this domain which requires transparent, open-source, and auditable technologies.

Difficulties in Attribution. A research direction that helps enable oversight, and evaluate and improve the correctness of solutions is self-consistency checks. Wang et al. (2023) used a prompting strategy that samples a diverse set of reasoning paths and then selects the most consistent answer. Fluri et al. (2024) proposed a framework to evaluate superhuman models by checking if they follow interpretable human rules, e.g., counterfactuals should flip the predicted decisions. Creating similar tests for OE AI is more difficult. One can change the parameters of the initial state of an OE AI to create a counterfactual environment; however, due to compounded cascading effects, the effects of the changed parameters cannot be easily isolated and are entangled with other novelty-related randomized intermediate states.

3.5 RESOURCE CONSTRAINTS

As the OE AI runs longer, it generates increasingly complex artifacts that require more computational and human resources to evaluate. Unlike traditional ML models, OE AI requires more continuous evaluation without clear guarantees of utility. OE AI is run for a longer time before producing useful results since it involves much exploration and is not targeted toward specific useful results. Furthermore, it is difficult to predict whether an OE AI will produce valuable artifacts. Thus, the significant computational resources might not be justified. These issues are exacerbated in OE AI that employs an LLM as a backbone since their large parameter size makes them expensive to run compared to smaller specialized models. Therefore, developing OE AI with adaptive resource constraints is important.

3.6 TRADE-OFFS

As the OE AI systems evolve, they must balance competing priorities, often resulting in trade-offs that make the deployment of these systems challenging. As explained in Figure 2, OE AI inherently faces a trade-off between speed, novelty, and safety, creating a trilemma where optimizing two of these dimensions often compromises the third. Speed refers to the rate at which the system can generate new artifacts. Novelty measures the degree of uniqueness or originality in each newly generated artifact. Safety represents the system's adherence to predefined constraints, ensuring outputs avoid harmful, unethical, or undesirable outcomes.

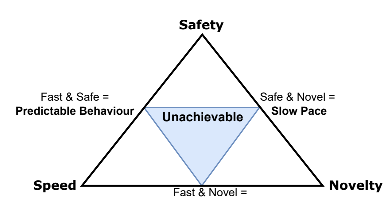

Figure 2: The Impossible Triangle of OE AI illustrates that safety, speed, and novelty cannot be maximized together.

Application-Specific Needs. Trade-offs can be difficult to navigate because they can depend on the types of problems we use OE AI for, which may require specific emphasis on one of these dimensions. In safety-critical applications such as drug discovery or medical diagnosis, safety is the foremost concern, often necessitating slower exploration to ensure rigorous validation and prevent harm, limiting novelty and speed. Conversely, in applications like gaming or art, novelty is prioritized to foster creativity, where the associated risks are generally lower, allowing safety to be sacrificed in favor of rapid, diverse output generation. Lastly, in autonomous vehicles or real-time industrial systems, the focus is on quick, reliable responses, with novelty being a secondary concern to ensure the system can operate effectively in dynamic, time-sensitive environments.

3.7 Social and Human risks

It is crucial to consider the societal risks of OE AI. While all new technologies may have negative societal consequences, the unpredictable and evolving nature of OE AI may amplify known AI harms or introduce unanticipated ones.

The Rate of Novelty. OE AI generates more novel artifacts than traditional AI and the rate of innovation and disruption is harder to anticipate. This might outpace society's ability to adapt, integrate, and understand new developments. History provides examples of the disruptive effects of excessive novelty, such as the Industrial Revolution, which, while transformative, led to widespread social upheaval, labor displacement, and the erosion of traditional ways of life. Purely AI-led innovation can result in a loss of human agency in shaping scientific and societal progress, leaving individuals feeling disconnected from the process of discovery and creation.

Uninteresting Artifacts. OE AI should produce results that are interesting and useful to the observer. Quantifying "interestingly new" progress has been one of the grand challenges in OE research. Foundation models have been used as a Model-of-Interestingness (MoI) Zhang et al. (2024b) to denote the human notion of what can be considered "novel" and at the same time "interesting". However, OE AI could still produce uninteresting artifacts. This could be because its sense of interestingness might be misaligned with ours or because it may get stuck in a narrow set of artifacts without exploring more widely. Also, as artifacts can be very complex it can be difficult for humans to determine whether they are truly interesting and useful. This could lead to situations where an OE AI produces useless, uninteresting artifacts, while humans do not recognize this. If such a system is kept running it will be a waste of resources. Furthermore, it might limit human creativity if human ideas are biased by generated artifacts or if humans think there is nothing more to explore. Such problems are now discussed with LLMs and how they can homogenize individuals' beliefs and lead to a false impression of consensus Burton et al. (2024).

Difficulty to Plan. As discussed in Section 3.5, it is intractable to foresee, plan, or track the OE AI's progress or whether it would produce valuable solutions. Given limited resources, we may need to prioritize which problems we delegate to OE AI. This has a resemblance to funding decisions for research proposals. Our society needs transparent, fairways of deciding on appropriate allocations.

Reshaping Human Values. LLMs may learn to mislead humans as a result of reward hacking Wen et al. (2024), wrongly convincing human evaluators that performance has increased. Persuasion, deception, or drifting to rogue goals are examples of the catastrophic risks of AI discussed in the literature with anticipation of becoming more likely when AI is adaptive Hendrycks et al. (2023), such as the case in OE AI. Due to cascading effects, OE AI may generate solutions that are initially, and then increasingly, misaligned, such as inaccurate scientific findings or biased policies. Values within societies may also drift over time, sometimes for the worse Hendrycks et al. (2023). As humans continue to get exposed to these proliferating artifacts, they might get normalized and set harmful precedence, i.e, OE AI may gradually change societal and human values instead of getting OE AI aligned to human values.

Accountability. Assigning accountability for the actions of traditional AI is an ongoing legal and ethical debate. However, it is even more complicated for OE systems, since they act autonomously and inherently behave in ways they were not designed to. This makes it unclear whether developers can be blamed for the wrongdoings of the model. Furthermore, OE AI does not follow traditional procedures for training and data collection, requiring new frameworks for assigning responsibility.

Environmental Factors. Current AI models use exuberant amounts of energy. Training GPT-3 consumed 1287 MWh of electricity, resulting in 502 metric tons of carbon emission Patterson et al. (2021). Data centers use around 2.5 percent of global electricity, rivaling the aviation industry in greenhouse gas emissions Pfeiffer (2023). The current paradigm of OE AI uses LLMs as a backbone; running these models continuously requires a high amount of computation, which can have a significant environmental impact.

4 Technical Mitigations of Risks

To address the risks and challenges, we explore and suggest research directions that enhance safety against catastrophic risks while responsibly maintaining the benefits of OE exploration.

4.1 Oversight

As it is hard to anticipate the safety of OE processes, it is critical to oversee, either by humans or another system, their behavior during execution. Oversight provides a mechanism to monitor, guide, and correct system behavior, ensuring outputs align with human values and safety expectations.

Human-in-the-Loop Oversight. Ultimately, only humans can define safety and desirable values. Thus it is critical to have a human in the loop when OE AI is run. This could mean that a human actively monitors new artifacts. The human overseer could intervene when unsafe artifacts are generated or filter which artifacts should be propagated to future iterations of the system. Furthermore, a human overseer could provide feedback and guidance that steers the OE process in interesting directions. OE can involve AI and human components working together Secretan et al. (2008). However, humans are limited in their capacity and might not be able to accurately judge complex artifacts, but should nevertheless set standards to remain in control.

Interpretable Decision-Making. To facilitate humans in providing oversight, future research needs to create interpretable OE AI whose decisions and reasoning traces are transparent to a human observer. Forcing OE AI to reason about its decisions in natural language, makes it inherently interpretable to humans, allowing inspection and failure detection Hu & Clune (2024); Betley et al. (2025). Systems can be trained to explain their artifacts to a weaker model to simulate a human overseer. Furthermore, interpretability tools can be used to understand which input features Wang et al. (2024b) or inner representations Alain & Bengio (2018); Cunningham et al. (2023) were relevant to a decision.

Hierarchical Oversight. Oversight can be expensive when a human or a large model needs to check every artifact. Hierarchical oversight can structure the supervision into layers, where a less expensive monitoring process oversees every artifact and reports artifacts or behaviors to higher levels with more expensive supervisors. Works such as Christiano et al. (2018); Chavan & Chavan (2024) propose mechanisms where higher layers guide or intervene in the functioning of lower layers. By analyzing the system's outputs at multiple levels of abstraction, hierarchical oversight can identify risks before they escalate while being resource efficient.

Scalable Oversight. Providing effective oversight is difficult for humans when generated artifacts become too complex for them to evaluate accurately. Scalable oversight seeks to align AI systems whose outputs surpass human expertise or are too numerous for humans to evaluate properly Burns et al. (2023). Approaches such as Iterated Distillation and Amplification (IDA) Christiano et al. (2018), Debate Irving et al. (2018) or Recursive Reward Modeling (RRM) Ibarz et al. (2018) could be applied to ensure the safety of OE AI. For example, OE AI could be forced to justify its actions in a debate with another agent, RRM could be used to align an overseer AI that can accurately evaluate new artifacts, or OE AI could be trained via IDA to internalize human notions of safety and interestingness. Furthermore, self-diagnostic tools such as Kamoi et al. (2024); Huang et al. (2024) can be applied to OE AI to detect vulnerabilities in the system.

OE AI for Adaptive Oversight. For OE AI, oversight should not only scale to complex artifacts but also accommodate the dynamic nature of OE AI by developing adaptable evaluation and uncertainty thresholds. An overseer needs to be able to generalize to novel, possibly OOD artifacts. OE AI itself can be used to develop new safety-specialized mechanisms that work in tandem with the diversity-driven OE AI. An example is an overseer OE AI that co-evolves and judges safety.

OE AI for Risk Extrapolation. Similarly, a specialized OE AI can be used to anticipate and simulate in advance the future trajectories of artifacts and assess their risks and cascading effects. This OE AI could be optimized to generate novel but specifically harmful artifacts. This would need quantification and uncertainty methods to measure how close the main artifact is to the hypothesized harmful ones, based on this, an abortion or intervention step can follow.

Consequential Actions. As OE AI continues to evolve and explore, it may intervene in its environments. We already observe strong progress in autonomous and embodied agents. However, for risky applications, e.g., scientific experiments, we would need to limit the OE AI from performing catastrophically consequential actions where we cannot yet anticipate their outcomes. An alternative is to build simulations and models that are faithful to our world that would enable sand-boxed artifact generation. Given the challenges posed by novel and emergent artifacts, exploring causal models is a promising direction, as they exhibit greater robustness on novel data Richens & Everitt (2024).

4.2 CONSTRAINTS

Most existing safety frameworks focus on structured environments with predefined goals. However, building guardrails to prevent the OE AI from exploring unsafe artifacts will be crucial to ensure the safety of these systems.

Constrained Exploration. Since OE AI often pursues diversity, the exploration process can inadvertently drive the system into unsafe or misaligned state spaces. By constraining exploration to an ϵ-ball, the system can balance novelty with safety, similar to safe exploration in RL Garcıa & Fernández (2015). This requires constrained novelty metrics that evaluate novelty relative to both past behaviors and predefined safety constraints. In simple, discrete domains, such a novelty metric could be formally specified, while LLM-based judges could quantify novelty in more complex domains. Based on the novelty scores of new artifacts, it would be possible to penalize novel behaviors that exceed a probabilistic safety threshold or confidence bound, as modeled using techniques like Gaussian Processes Sui et al. (2015); Turchetta et al. (2016) or reachability analysis Krakovna et al. (2018); Fisac et al. (2018). Furthermore, novelty search can be combined with shielding mechanisms Dawood et al. (2024) to dynamically reject unsafe actions. Finally, safety constraints also can be introduced in Minimal Criterion Coevolution Brant & Stanley (2017).

Artifact Complexity Budget. Setting a complexity budget might help balance novelty and exploration with the ability of humans to understand, evaluate, and digest new artifacts. This budget serves as a safeguard, preventing excessive unpredictability and mitigating the risk of negative compounding effects that may arise from unrestrained exploration. By dynamically adjusting this budget it is possible to navigate the creativity-control trade-off.

Setting Specific Rules. While OE AI continuously evolves and faces new challenges, there are rules we never want it to break. Although such rules cannot cover all unsafe behaviors, they can still prevent some failures. While constraints do limit the creativity of the OE AI by cutting off some of the search space, the system is still able to openly explore the remaining space, thus retaining its open-endedness. To take a more abstract and flexible view, rules could be specified as general principles in a constitution Bai et al. (2022) that can be reinterpreted in new situations, or dynamically created and updated by AI. An LLM guiding the OE AI's decisions can either reason about these rules Guan et al. (2025) or causally Kıcıman et al. (2024). Recent work Zaremba et al. (2025) shows the potential and promise of LLMs, when given enough intermediate reasoning steps, to reason in compliance tasks. This also provides an effective framework for overseers to judge new artifacts.

4.3 ADAPTIVE ALIGNMENT

Current alignment techniques assume a model and its environment remain static, thus only requiring safety training once. New continual alignment algorithms could allow us to adapt safety as the model and its circumstances change Zhang et al. (2024a). While Moskovitz et al. (2024) composite reward weighting dynamically and Hong et al. (2024) address overoptimization and ambiguity, they lack robust mechanisms for long-term feedback loops. Multi-agent RL for co-evolving alignment dynamics in OE systems can be a promising research direction. Using dynamic reward functions can adjust the reward signals to reflect the evolving human preferences or system performance. Adaptive preference scaling Fang et al. (2024); Hong et al. (2024), and distributional preference reward modeling Li et al. (2024) have been used to refine reward functions in RL-based systems by adjusting reward weights in response to shifting human feedback or performance degradation. For OE AI, dynamic reward calibration must go beyond simple reward adjustments to handle the continuous and diverse outputs produced by such systems.

4.4 SAFETY EVALUATIONS

Finally, continuous safety evaluation of OE AI is important for understanding the extent of unsafe behaviors.

Benchmarking OE Safety. Developing benchmarks specifically for OE AI is crucial for quantifying its risks and evaluating failure modes. Existing benchmarks, such as those on multi-agent risks and unintended consequences Rivera et al. (2020), provide some insights but fail to incorporate the unique characteristics of OE algorithms. A dynamic benchmark explicitly designed for OE AI would need to address its continuous evolution, novelty generation, and dynamic complexity. For example, the difficulty of tests could be adjusted to the OE AI's changing capabilities.

Redteaming OE Systems. The previously outlined direction of "extrapolating risks" is beneficial to anticipate future risks even if the OE system is aligned. On the other hand, targeted red teaming can reveal failures for individual components or the entire system. Red teaming allows us to stress-test OE systems by actively probing their vulnerabilities and finding situations in which they behave unsafely.This could involve manually or adversarially finding inputs on which the OE system misbehaves. Lehman et al. (2023); Bradley et al. (2023); Liu et al. (2024) uses LLMs to enhance genetic programming by generating diverse, functional artifacts. These outputs could serve as adversarial artifacts to test and evaluate system robustness like in Samvelyan et al. (2024), but here the aim would be to test the entire OE systems. Further, one could construct an environment in which the OE system is being led to produce unsafe artifacts.

5 Call for Action

Ensuring the responsible deployment of OE AI requires active engagement from various stakeholders.

Funding bodies can shape research priorities. They could urge OE researchers to consider and address the safety risks of their work. Further, they could dedicate resources toward robust safety mechanisms and evaluations for OE AI.

Research on the intersection of safety and OE research is crucial, impactful and under-explored. We argue that safety should be a critical part of OE research. This requires general awareness of the risks and dedicated research on safety problems. Additionally, the AI safety community should dedicate research to the specific risks of OE AI. We hope this paper can provide a bridge to foster exchange and collaboration between these communities.

Opportunities lie in the application of OE AI to AI Safety. Aside from providing adaptive oversight (Section 4.1) OE AI can be used to red-team traditional models Samvelyan et al. (2024) and agentic applications, in addition to automating interpretability research.

Policy Makers should mandate audits of sufficiently capable OE AI to ensure adherence to safety standards and societal values. Comprehensive auditing protocols must account for the dynamic and emergent nature of these systems.

Industry deploying OE AI must implement and rigorously test oversight mechanisms and guardrails for OE systems. Furthermore, comprehensive evaluation of societal and catastrophic risks should be conducted in collaboration with third-parties, academia and governments.

Public. The ability and resources to run OE AIs are centralized in a few companies. Since deploying them comes with large resource costs and safety risks, the public should be educated and consulted on these decisions to prioritize.

6 Conclusion

Open-Ended AI is a promising paradigm for generating novel, adaptive solutions in complex and dynamic environments, driving interest across research and applied domains. However, its open-ended nature introduces specific safety challenges that must be addressed to enable responsible deployment and maximize its societal benefits. We argue that the inherent unpredictability and uncontrollability of OE AI, challenges in ensuring and maintaining alignment, traceability, and societal impacts, as well as trade-offs in resource use and safety. We highlight the critical importance of human and automated oversight over OE AI. Further, we suggest ways of giving adaptive guidelines to OE AI that retain its creativity and co-evolve with it. Lastly, we call for targeted safety evaluations and provide concrete suggestions on how different stakeholders can contribute to the responsible development of OE AI. Ultimately, we hope this paper will lead the OE and safety communities and other stakeholders to consider safety a priority in the development and deployment of OE AI.

References

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

Fuma Aki, Riku Ikeda, Takumi Saito, Ciaran Regan, and Mizuki Oka. Llm-poet: Evolving complex environments using large language models. In *the Genetic and Evolutionary Computation Conference Companion*, 2024.

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, pp. 1–10, 2025.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv*, 2018.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: Llms are aware of their learned behaviors. *arXiv*, 2025.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv*, 2021.

Herbie Bradley, Andrew Dai, Hannah Benita Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Gregory Schott, and Joel Lehman. Quality-diversity through AI feedback. In *Second Agent Learning in Open-Endedness Workshop*, 2023.

Jonathan C Brant and Kenneth O Stanley. Minimal criterion coevolution: a new approach to open-ended search. In *the Genetic and Evolutionary Computation Conference*, 2017.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. [LINK], 2024.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv*, 2023.

Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, et al. How large language models can reshape collective intelligence. *Nature human behaviour*, 8(9): 1643–1655, 2024.

Emilio Cartoni, Davide Montella, Jochen Triesch, and Gianluca Baldassarre. Real-x–robot open-ended autonomous learning architectures: Achieving truly end-to-end sensorimotor autonomous learning systems. *arXiv*, 2020.

Emilio Cartoni, Davide Montella, Jochen Triesch, and Gianluca Baldassarre. Real-x—robot open-ended autonomous learning architecture: Building truly end-to-end sensorimotor autonomous learning systems. *Transactions on Cognitive and Developmental Systems*, 15(4):2014–2030, 2023.

Parikshit Chavan and Peeyusha Chavan. Automation of ad-ohc dashbord and monitoring of cloud resources using genrative ai to reduce costing and enhance performance. In *the IEEE International Conference on Innovations and Challenges in Emerging Technologies (ICICET)*, 2024.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv*, 2024.

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

Jeff Clune. Ai-gas: Ai-generating algorithms, an alternate paradigm for producing general artificial intelligence. *arXiv*, 2019.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv*, 2023.

Murad Dawood, Ahmed Shokry, and Maren Bennewitz. A dynamic safety shield for safe and efficient reinforcement learning of navigation tasks. *arXiv*, 2024.

Tej Deep Pala, Vernon YH Toh, Rishabh Bhardwaj, and Soujanya Poria. Ferret: Faster and effective automated red teaming with reward-based scoring technique. *arXiv*, 2024.

Aaron Dharna, Amy K. Hoover, J. Togelius, and L. Soros. Transfer dynamics in emergent evolutionary curricula, 2022.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *ICML*, 2022.

Adrien Ecoffet, Jeff Clune, and Joel Lehman. Open questions in creating safe open-ended ai: tensions between control and creativity. In *Artificial Life Conference Proceedings 32*, pp. 27–35. MIT Press, 2020.

Feiteng Fang, Liang Zhu, Xi Feng, Jinchang Hou, Qixuan Zhao, Chengming Li, Xiping Hu, Ruifeng Xu, and Min Yang. Clha: A simple yet effective contrastive learning framework for human alignment. In *the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *Transactions on Automatic Control*, 64(7):2737–2752, 2018.

Manon Flageat and Antoine Cully. Uncertain quality-diversity: evaluation methodology and new methods for quality-diversity in uncertain domains. *Transactions on Evolutionary Computation*, 2023.

Manon Flageat, Hannah Janmohamed, Bryan Lim, and Antoine Cully. Exploring the performance-reproducibility trade-off in quality-diversity. *arXiv*, 2024.

Lukas Fluri, Daniel Paleka, and Florian Tramèr. Evaluating superhuman models with consistency checks. In *SaTML*, 2024.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models, 2025.

Vernon Toh Yan Han, Rishabh Bhardwaj, and Soujanya Poria. Ruby teaming: Improving quality diversity search with memory for automated red teaming. *arXiv*, 2024.

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

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic ai risks. *arXiv*, 2023.

Ilgee Hong, Zichong Li, Alexander Bukharin, Yixiao Li, Haoming Jiang, Tianbao Yang, and Tuo Zhao. Adaptive preference scaling for reinforcement learning with human feedback. In *NeurPS*, 2024.

Shengran Hu and Jeff Clune. Thought cloning: Learning to think while acting by imitating human thinking. *NeurIPS*, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *ICLR*, 2024.

Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktaschel. Open-endedness is essential for artificial superhuman intelligence. *ICML*, 2024.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. In *NeurIPS*, 2018.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv*, 2018.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2024.

Minqi Jiang, Tim Rocktäschel, and Edward Grefenstette. General intelligence requires rethinking exploration. *Royal Society Open Science*, 10(6):230539, 2023.

P. Kafka. The automotive standard iso 26262, the innovative driver for enhanced safety assessment & technology for motor cars. *Procedia Engineering*, 45:2–10, 2012.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. In *ICML*, 2024.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *TMLR*, 2024.

Victoria Krakovna, Laurent Orseau, Ramana Kumar, Miljan Martic, and Shane Legg. Penalizing side effects using stepwise relative reachability. *arXiv*, 2018.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity. [LINK], 2020.

Joel Lehman and Kenneth O Stanley. Revising the evolutionary computation abstraction: minimal criteria novelty search. In *the 12th annual conference on Genetic and evolutionary computation*, 2010.

Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

Joel Lehman, Jonathan Gordon, Shawn Jain, Kamal Ndousse, Cathy Yeh, and Kenneth O Stanley. Evolution through large models. In *Handbook of Evolutionary Machine Learning*, pp. 331–366. Springer, 2023.

Nancy G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, 2012.

Dexun Li, Cong Zhang, Kuicai Dong, Derrick Goh Xin Deik, Ruiming Tang, and Yong Liu. Aligning crowd feedback via distributional preference reward modeling. *arXiv*, 2024.

Shengcai Liu, Caishun Chen, Xinghua Qu, Ke Tang, and Yew-Soon Ong. Large language models as evolutionary optimizers. In *2024 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8. IEEE, 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv*, 2024.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Position: Levels of agi for operationalizing progress on the path to agi. In *ICML*, 2024.

Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca Dragan, and Stephen Marcus McAleer. Confronting reward model overoptimization with constrained RLHF. In *ICLR*, 2024.

Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv*, 2015.

Eleni Nisioti, Claire Glanois, Elias Najarro, Andrew Dai, Elliot Meyerson, Joachim Winther Pedersen, Laetitia Teodorescu, Conor F Hayes, Shyam Sudhakaran, and Sebastian Risi. From text to life: On the reciprocal relationship between artificial life and large language models. In *Artificial Life Conference Proceedings 36*, volume 2024, pp. 39. MIT Press, 2024.

Norman Packard, Mark A Bedau, Alastair Channon, Takashi Ikegami, Steen Rasmussen, Kenneth O Stanley, and Tim Taylor. An overview of open-ended evolution: Editorial introduction to the open-ended evolution ii special issue. *Artificial life*, 25(2):93–103, 2019.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv*, 2021.

Eric Pfeiffer. Wired - the true cost of generative ai: Data centers and energy consumption. [LINK], 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. *ICLR*, 2024.

Corban G Rivera, Olivia Lyons, Arielle Summitt, Ayman Fatima, Ji Pak, William Shao, Robert Chalmers, Aryeh Englander, Edward W Staley, I Wang, et al. Tanksworld: a multi-agent environment for ai safety research. *arXiv*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H. Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Nicolaus Foerster, Tim Rocktäschel, and Roberta Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts. In *NeurIPS*, 2024.

Jimmy Secretan, Nicholas Beato, David B D Ambrosio, Adelein Rodriguez, Adam Campbell, and Kenneth O Stanley. Picbreeder: evolving pictures collaboratively online. In *the SIGCHI conference on human factors in computing systems*, 2008.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv*, 2022.

Olivier Sigaud, Gianluca Baldassarre, Cédric Colas, Stephane Doncieux, Richard Duro, Pierre-Yves Oudeyer, Nicolas Perrin-Gilbert, and Vieri Giuliano Santucci. A definition of open-ended learning problems for goal-conditioned agents. *arXiv*, 2023.

Lisa Soros and Kenneth Stanley. Identifying necessary conditions for open-ended evolution through the artificial life world of chromaria. In *Artificial Life Conference Proceedings*, pp. 793–800. MIT Press, 2014.

Lisa B Soros, Joel Lehman, and Kenneth O Stanley. Open-endedness: The last grand challenge you've never heard of, 2017.

Kenneth O Stanley. Why open-endedness matters. *Artificial life*, 25(3):232–235, 2019.

Kenneth O Stanley and Joel Lehman. *Why greatness cannot be planned: The myth of the objective*. Springer, 2015.

Kenneth O Stanley and L Soros. The role of subjectivity in the evaluation of open-endedness. In *Presentation delivered in OEE2: The Second Workshop on Open-Ended Evolution, at ALIFE 2016*, 2016.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *ICML*, 2015.

K. Suyama. Probabilistic safety assessment and management of control laws. In *the 35th Annual Conference of IEEE Industrial Electronics*, 2005.

Open Ended Learning Team, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, et al. Open-ended learning leads to generally capable agents. *arXiv*, 2021.

Matteo Turchetta, Felix Berkenkamp, and Andreas Krause. Safe exploration in finite markov decision processes with gaussian processes. *NeurIPS*, 2016.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024a.

Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O Stanley. Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions. *arXiv*, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. Gradient based feature attribution in explainable ai: A technical review, 2024b. URL `https://arxiv.org/abs/2403.10415`.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R Bowman, He He, and Shi Feng. Language models learn to mislead humans via rlhf. *arXiv*, 2024.

Marvin Zammit, Antonios Liapis, and Georgios N Yannakakis. Map-elites with transverse assessment for multimodal problems in creative domains. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*. Springer, 2024.

Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, and Johannes Heidecke Amelia Glaese. Trading inference-time compute for adversarial robustness. 2025.

Han Zhang, Yu Lei, Lin Gui, Min Yang, Yulan He, Hui Wang, and Ruifeng Xu. Cppo: Continual learning for reinforcement learning with human feedback. In *ICLR*, 2024a.

Jenny Zhang, Joel Lehman, Kenneth Stanley, and Jeff Clune. OMNI: Open-endedness via models of human notions of interestingness. In *ICLR*, 2024b.